

AI-assisted cervical cytology precancerous screening for high-risk population in resource-limited regions using compact microscope

Corresponding Author: Dr Xiuli Liu

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

Review report

The authors present a cervical cytology precancerous screening system specifically designed for high-risk populations in resource-limited regions. This system integrates a cost-efficient and compact microscope with an AI algorithm to identify HSIL cases. The proposed solution aims to address the bottlenecks in conventional cervical cancer screening, particularly the high costs and reliance on qualified pathologists. The screening system training employs a two-stage slide classification model: 1. The first stage reduces negative cells through self-supervised pretraining, 2. The second stage uses an attention-based transformer (Attention-Transformer) to classify slides by aggregating deep features from instances with the highest lesion probabilities. The system has been validated on a multi-center dataset, demonstrating promising results with high AUC values for screening high-risk SIL cases. This approach has the potential to improve cervical cancer screening coverage in underdeveloped regions, contributing to global health initiatives. Besides the above merits, we have the following points.

Major point

- The development of a low-cost, compact microscope suitable for resource-limited areas, with an approximate pixel resolution of 0.87 $\mu\text{m}/\text{pixel}$ in line 103. I am not sure if the resolution is sufficient to make the diagnosis of other types of cells, for example, LSIL, ASCUS, etc. Or if this is the actual limitation that the system can only use for the high-risk squamous intraepithelial lesion (HSIL).
- The author introduced a dual-stage slide classification model that effectively handles sparsely-distributed lesion cells in cytology whole slide images (WSIs), it is different from the common learning methods, using the supervised for initial training and then apply self-supervised learning to boost the model performance, probably unsupervised learning as last step. Vision Transformers (ViTs) typically demand substantial amounts of training data to achieve competitive performance levels. The original study (Dosovitskiy et al., "An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale," ICLR 2021) concluded that ViTs outperform the top ResNet models across all public datasets, contingent on sufficiently large pre-training datasets. This advantage becomes more pronounced with increasing dataset sizes. When leveraging ViTs with sufficiently large datasets (even for challenging tasks like identifying lesion cells), their performance excels. However, with smaller datasets, ViTs may exhibit slightly lower performance compared to other classification models such as ResNet. The authors need more explain the rationale of the use of self-supervised pretraining on a large number of unlabelled cervical images first, then followed by transfer learning on a smaller set of labeled images.
- The explanation of the Attention Transformer's specific modifications compared to traditional transformers is brief. More detailed architectural insights would help us understand its unique contributions.
- The authors mentioned the implementation of an attention-based transformer for aggregating deep features from high-probability lesion instances to improve slide classification accuracy, while the feature extract is not well reported in the manuscripts. BTW, we understand that it is often that the cells may form clusters which may bring another challenge to the accurate diagnosis. Is this scenario addressed using the same set of features? Or such clusters were filtered out by the model?
- In line 109, the authors state, "Deep features extracted from redundant instances with the top 200 lesion probabilities are aggregated." Could the authors clarify why the top 200 was chosen specifically? Why not the top 100, top 300, or another number? Considering there are tens of thousands of cells in the whole slide images (WSI) of cervical cytology as mentioned

in line 79, the rationale behind selecting this specific number would be valuable for understanding the methodology. The paper does not sufficiently detail the criteria for selecting the top 200 instances with the highest lesion probabilities. Clarification on this selection process would enhance reproducibility.

- While the final performance metrics are provided, there is a lack of discussion on potential false positives and false negatives and their implications on clinical decisions. In addition, some important intermedia results will also be appreciated if included in supplementary material.
- Authors acknowledge the variability in slide preparation and staining processes across different regions, which can affect AI algorithm generalization. The visual comparison was provided in Figure 1b and a supp figure, while a quantitative comparison will be better to understand the difference between the images and how they impact the diagnosis. For example, the authors mention providing 100 pairs of images for an imaging quality check in line 153. It would be beneficial to include quantitative measurements, such as the percentage of lost focus, the frequency of saturation, noise levels, etc., for a more comprehensive evaluation of the images generated by the compact microscope. Currently, a detailed assessment of the generated images is not provided.
- Although the Attention Transformer provides some level of interpretability through attention scores, further efforts to enhance model transparency and explainability would be beneficial. Techniques such as saliency maps or other visualization tools could be employed to provide clearer insights into the model's decision-making process.
- The study successfully uses low-resolution WSIs, but there are inherent limitations in detecting fine details, such as low-grade lesions or atypical cells, which is sometimes more important for early detection. Not sure what is the performance of those cell types.
- The manuscript would benefit from the expertise of a professional editor to enhance the language. Currently, it is challenging to read, piecewise lacks detailed information, and is not well organized. Additionally, the supplementary documentation does not adequately aid in understanding the experiment and implementation.
- Some recent publications on this topic as follows are not cited, it is also important to differentiate this work from the existing works:

1. Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears
2. Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer

Minor points

1. In line 184, "The right figure shows the average pixel intensity", it is clear that the y-axis represents pixel grey intensity. What is the x-axis? Micro or pixel?
2. In line 196, citation number 40 cannot be found in the manuscript, as only 37 citations are listed. Please check all citations and references.
3. Could you please clarify the difference between cell level and instance level in Figure 2? Are they referring to the same concept? Please consider adding the dimensions of the patches (224x224 and 256x256) in the figure for clarity. Ensuring that the main text and the figure complement each other will aid the reader's understanding. Currently, it is unclear which is the first and which is the second stage in the figure. While the figure is presented, it is not helpful.
4. In line 483, Camelyon16 reported a dataset comprising 400 WSIs. The first training dataset includes 170 WSIs of lymph nodes (100 normal and 70 containing metastases), while the second consists of 100 WSIs (60 normal and 40 containing metastases). The test dataset comprises 130 WSIs collected from both universities. However, there is uncertainty regarding the standard histology of 110 WSIs and the criteria for their selection.
5. In line 484, The process of creating 106 fully annotated standard WSIs is unclear. Specifically, details such as the number of cytologists involved, and any cross-checks and validation procedures for the annotated data, are not specified.
6. In line 491, the process from 224 pixels to obtain 16 pixels in this study is unclear; the sentence appears to be incomplete.
7. Could you please provide details regarding the accuracy in detecting negative and challenging samples described in line 522? Additionally, could you elaborate on the validation process used to identify these challenging samples?
8. In line 561, upon reviewing the references cited in this sentence, I couldn't locate any related sources, including supplementary materials, regarding this augmentation process. I kindly suggest considering the addition of more detailed information to this manuscript.
9. Improving the formatting of subsections would be beneficial, as exemplified in lines 463, 480, 527, 536, and so forth.

The paper presents a promising approach to cervical cytology precancerous screening in resource-limited regions, leveraging a low-cost compact microscope and an AI-based classification system. While the results are encouraging, some above issues need to be well addressed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

This paper developed a cervical precancerous assisted-screening system for identifying high-risk squamous intraepithelial lesion cases, which was composed of a low-cost miniature microscope and a low-pathologist-reliance artificial intelligence algorithm. Though the research is interesting and may have a wide clinical application, the novelty and the presentation of this work is still limited. In addition, I have following other concerns:

1. The abstract part fails to reflect the novelty of the paper. Please further describe the hardware and software innovations of the developed system in detail.
2. While AI-assisted cancer screening has been successfully applied in several related fields, the authors need to show at least killer applications for the demonstration of their method with demonstration of their unique advantages. Currently, the demonstrations can be achieved by many previous methods.
3. Please add what is the unit of the horizontal coordinate on the right picture of Figure 1 b.
4. It seems that there is a contradiction between the sentence "In computational pathology field, there have been many high-performance weakly-supervised WSI classification methods which are built without the need for cell-level annotations. However, there are few such kind of methods for cytology WSIs" and "On the contrary, most of them need huge number of cell-level annotations as mentioned in the introduction". Are cell-level annotations needed or not for the mentioned methods?
5. The paper repeatedly mentioned that the poor performance of the comparison methods was due to the "limited resolution of our compact microscope" and "sparsely-distribution lesion cells". Please add the corresponding results in the text to show the influence of the above two reasons on the comparison method.
6. I want to know why the accuracy of Figure 5 and Figure 4 is different, I mean if I understand correctly, the results of the "supervised" in Figure 5 should be the same as "ours" in Figure 4, but they are different in the paper.
7. The manuscript needs to be polished severely. Plenty of grammar errors can be found.

(Remarks on code availability)

The codes are fine with references.

Reviewer #5

(Remarks to the Author)

Bai et al. demonstrated a low-cost compact microscope and an AI classification algorithm for cervical cytology to enable cancer screening in resource-limited settings. The prototype microscope captured similar cell morphology features compared to the standard microscope but with reduced cost and size. The authors developed a two-stage classification algorithm based on a Transformer model, which was then trained and validated using microscopic images of cytology slides captured by the proposed compact microscope. However, I cannot recommend this manuscript to be published in Nature Communications due to several significant drawbacks in this paper. To validate the clinical performance of the proposed method, a more comprehensive analysis of the testing accuracy is necessary. In addition, significant improvements to the system setup and imaging procedure are also required to facilitate its application in various clinical settings, especially in resource-limited areas. My comments are listed below.

Major comments:

1. The authors evaluated the performance of the proposed method and the screening accuracy is shown in Figure 4. The authors make the argument that the model can generalize well across datasets. However, when testing the system using the two external datasets and NewCohorts dataset, the authors exclude atypical squamous cells (ASC) cases because of the lower classification performance. ASC is the most common abnormal finding in cervical cytology, which can also be reflected in the number of ASC cases within the patient cohort. In addition, ASC cases were included in the training dataset. Given these reasons, the authors should include ASC cases in all cases when evaluating the system performance, including Figs. 4, 5 and 7. The sensitivity and specificity considering all positive cases should be evaluated and reported. The authors should further optimize the hardware or the software of the proposed system if the accuracy is compromised.
2. As shown in Fig. 4, the sensitivity and specificity of the proposed method varied a lot when using different datasets. Could the authors comment on that?
3. How are the classification metrics calculated? Is this done at WSI level or individual tile level? How do you account for the redundant images? The authors should detail this in the methods section.
4. Although the authors proposed a compact microscope that can be used in resource limited settings, only a rough, incomplete system prototype was demonstrated in this study, which still requires significant optimization to improve usability of the device in the clinic. The image acquisition relied on manual movement of the slides and capturing around 90 images to cover the entire field of view, which is highly labor-intensive and unreliable. The authors should automate the image acquisition and analysis procedure to enable the clinical application of the system, or the authors should acknowledge this limitation in the manuscript. The system packaging should also be further optimized. It would be helpful to include a clear description or flow chart detailing which portions of the image acquisition are manual and which are automated.
5. Compared to a standard microscope, the images acquired using the proposed compact microscope exhibit a different color tone with a pronounced purple/blue background across the entire field of view. Could the authors provide an explanation for this? Has any white balance correction been explored? While we understand that the algorithm may be able

to compensate for this if it is only trained on images from this microscope, any real world implementation of the system should allow for the visual review of images by a pathologist when a pathologist is available, and the while we understand there may be some artifacts due to the low-cost nature of the microscope, it seems that a white balance correction should be implemented.

6. The authors demonstrate a lower contrast and higher noise level of the proposed compact microscope compared to the standard microscope. However, instead of qualitative comparisons of the image quality of the 100 image pairs, the authors should provide a comprehensive quantitative analysis to compare the two imaging systems.

Minor comments:

1. The comparisons between cytology images and histopathology images shown in Fig. 1c and 1d are distinct from the other panels in Fig. 1. The authors might consider moving Fig. 1c and 1d to a different figure or supplementary materials. Consider removing Figure 1d entirely, as it does not contribute to the focus of your paper.

2. For the input of the first stage of the network, why is the image patch resized to a smaller size?

3. When dividing a large image to smaller 256*256 image patches, what is the size of the overlapped area between two neighboring patches? Could the authors explain why a redundant sampling is needed, and how is the size of the overlapped area determined?

4. How many patients are included for each dataset used in this study?

5. It is unclear whether inference was run in a real-time or near real-time manner with images collected with the microscope system. What are the barriers here?

6. The paper should be proofread to correct many grammatical errors, typos, and general edits. A few are noted below, however this is not a comprehensive list:

- a. Define what negative and positive cases are in the introduction.
- b. Typo in "sparsely-distributed lesion cells" found throughout the manuscript.
- c. In Figure 7, the legend should be "HSIL" instead of "HISL". Also CLAM is mentioned twice in the caption.
- d. In the abstract, "for high-risk population" should read "for high-risk populations"
- e. On page 3, "easy of connection" should read "ease of connection"

7. It would be helpful to consistently refer to the new method as either "OUR" or "Att-Transformer" in the figures. The use of "OUR" is somewhat confusing as it can appear to be an undefined acronym.

8. The Figure 3 caption needs to be expanded to include additional details. For example, an explanation of the term "hard samples" would be helpful.

9. The term "grassroots" is a bit unorthodox to describe a low resource setting. The authors may want to reconsider this term.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)

This is a manuscript investigating AI supported cervical cytology for intended use in low-resource settings. Whereas the goal of the research is highly commendable, and the work is very timely and interesting, I have the following major comments for consideration.

The Introduction states in lines 49-57, that "cytology screening remains crucial for reducing the incidence of cervical cancer". This completely ignores the possibility of HPV-based screening, with or without the addition of self-sampling in a low-cost medium, which is a high throughput technology at low cost. Substantial efforts worldwide have been dedicated to push the cost of cervical screening down through clever use of HPV testing to rule out women with no risk for cancer. And it is surprising to see no acknowledgment or position from the authors on this possibility, especially as they allude to the WHO elimination goal on row 134.

The problem of cervical cytology is that of subjectivity in diagnosis and particularly, false negatives, and correct training of the algorithm is essential, as well as maintenance. The European Union will shortly advise against the use of cytology instead of focusing on the importance of introducing HPV-screening for the good of the female population. I am surprised at the lack of consideration in the paper on how the authors defined and validated the gold standard diagnoses of subclasses ASC, LSIL and HSIL, especially since ASC suffers from notorious inter-observer variation. (AI could hopefully improve substantially upon this situation, indeed, but right now I need to understand better how the algorithm was basically validated as correct.)

That said, the large remaining problem is how to handle triage of HPV-positive women to the limited resources of

colposcopy and biopsy. Here, cytology of HPV-positive smears with AI-assisted efforts could be useful and I was somewhat surprised to see that this was not the focus of the study. For Nature group to consider this manuscript further, it is my opinion that the authors must conduct a substudy in HPV-positive samples to present a more timely case compatible with present global algorithms for more regions than those in China.

I entirely miss reports on which set of pathologists/panels ultimately classified and validated what constituted an abnormal cell, or a true lesion apparently defined as “subclasses ASC, HSIL, LSIL, NEG” – ASC and LSIL in particular are notoriously subjective reference standard issues. Row 454 should state ASC, LSIL, HSIL, NEG, by the way.

It is not surprising that the AUC is 0.89 on NewCohorts as the algorithm essentially returns to the original 4 centres' type of classification of subclasses – is there some form of circular over-fitting going on here? Overall I find an AUC of 0.90 to be very optimistic, especially given that cytology diagnosis of underlying histological lesions is not a very precise science to begin with.

Ideally, a fully external evaluation of the tiny microscope and classifier should be done by a collaborating partner in a clinic/pathology laboratory reference cytology of very high quality, for example in the US or Europe.

Minor comments:

I went to the authors' github account to try to look at the Demonstration data alluded to in row 580 but could not immediately locate that file. Please clarify exactly where in the github material this is located.

“Grassroots centers” – what is meant by this term exactly?

Row 286-287 typos, it should be HSIL and LSIL.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The revised manuscript has been substantially improved and presents a promising approach to cervical cytology precancerous screening in resource-limited regions by integrating a compact, cost-effective microscope with an AI-based classification system. The authors have provided detailed explanations in response to the reviewers' comments.

The proposed two-stage classification model, which combines self-supervised learning with an attention-based transformer, is well-motivated and supported by small-scale multi-center validation. While the results demonstrate strong potential, several areas require minor further clarification and improvement.

One previous key concern raised previously was the resolution of the compact microscope, stated to be approximately 0.87 $\mu\text{m}/\text{pixel}$. The authors have addressed this issue by including visualized images and quantitative data.

I understand the rationale behind the two-step training approach, particularly with the support of the reply to the review comments. However, since readers may not have access to these responses, I suggest that the authors further elaborate on this idea within the manuscript.

The robustness of the model is discussed based on the given dataset, but additional field testing and validation may be necessary in the future, which may not be within the scope of this work.

The manuscript would benefit from professional editing to enhance readability and organization. Ensuring consistency between figures, text, and citations (e.g., missing references in citation lists) is essential for completeness.

Overall, this study presents an innovative AI-assisted cervical screening system with conceptual advancements aimed at addressing healthcare challenges in regions with limited medical resources. Additionally, some of the content from the “reply to reviewer” document provides valuable insights that could help readers better understand and appreciate this work. The authors may consider making their responses more concise and incorporating key points into the main text.

Recommendation: Accepted. Further review is not necessary.

Other Minor Issues:

- Figure quality needs improvement. Some figures could be reorganized and combined for clarity. The figure style should also be adjusted to align with Nature journal standards. Authors may also consider to improve the quality of the figure captions.
- Ensure consistent abbreviation of author names in the author contributions section. For example, Liu Xiuli → X.L.L., but Shenghua Cheng → S.C. Please review this section for consistency and clarity.
- References [39] and [42] are duplicated. Please check the full reference list to eliminate such errors.
- All acronyms should be defined when first introduced in the text. For instance, “SoC” appears in line 180, but its full name

is only explained in line 548. Please ensure this principle is applied consistently throughout the manuscript.

Review Comments from the Junior Reviewer:

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

This study introduces a low-cost, compact microscope combined with an AI-based screening system to enhance cervical cytology in resource-limited settings. By employing a two-stage approach—instance classification followed by an Att-Transformer model—the system achieved AUCs of 0.85–0.89 across multiple datasets, demonstrating its potential for efficient screening with minimal resources. The manuscript has improved significantly, with the authors effectively addressing my concerns and incorporating substantial experiments to support their conclusions. The reorganization enhances readability, and the supplementary materials provide valuable additional context. Overall, I appreciate the improvements and found the manuscript engaging to read.

Minor comments

In Figure 1b, AB is labeled as the standard microscope in the plot. However, I noticed that the AB line is positioned in the lower-left section, which corresponds to the LR image acquired by the compact microscope. Please revise accordingly.

(Remarks on code availability)

I did not assess the code as I am currently on a trip.

Reviewer #2

(Remarks to the Author)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Reviewer #4

(Remarks to the Author)

The authors have fully addressed my concerns during the revision with additional experiments.

(Remarks on code availability)

The code is fine with a readme file.

Reviewer #5

(Remarks to the Author)

The revised manuscript states “Cytopathology-based screening, 72 which directly identifies cellular morphological changes, remains a necessary follow-up step after 73 a positive HPV test for definitive diagnosis of cervical precancer.” This is not correct. A positive HPV test is followed by colposcopy and biopsy with histopathologic diagnosis of cervical precancer. The incorrect statement added to the introduction implies that the new methodology can be used for definitive diagnosis. This is not consistent with accepted clinical practice.

The revised paper does provide AUC for datasets including ASCUS. However, performance suffers quite a bit and these results are presented in supplementary figures. It is not clear that the performance achieved with benign samples included is clinically relevant even if the method does outperform other computational approaches. The authors correctly state that it is important to detect HSIL lesions, but they must still be discriminated from ASC samples. “Our method achieved overall average AUC of 0.780 whereas AUCs of the comparison 334 methods are lower than 0.525 (supplementary Fig. 7).”

Finally, recommendations for co-testing with HPV and Pap generally recommend that patients are referred for colposcopy if the HPV test is positive even if the Pap is negative.

(Remarks on code availability)

Reviewer #6

(Remarks to the Author)

I find the responses to my previous comments satisfactory. My two remaining comments are rather regarding the response letter. Firstly, in the response to my initial comments the authors state that HPV screening is appropriate for “lower-resource” regions. It is important and appropriate, as well as much better than cytology, in any resource setting. Cytology is no longer the primary option even in higher-resource settings. Secondly, the authors on page 30 of the response letter state “In combined screening, our model can effectively exclude a large number of cytopathology negative cases, thereby improving

the specificity of co-testing. For instance, we can recommend patients who are HPV-positive and have the highest scores of our algorithm for biopsy, while those with relatively lower algorithm scores but still HPV-positive can be referred for endoscopic examination. "

We do not really use the term "endoscopy" for visualizing the cervix. Do you mean colposcopy? Biopsies are taken under colposcopy guidance, so the above sentence is very difficult to understand - group 1 will need referral for colposcopic examination if a biopsy is to result.

(Remarks on code availability)

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to reviewers

We thank the editor and the reviewers for their constructive critiques, which have been extremely helpful in our effort to improve the manuscript. In the revised manuscript submitted here, we have attempted to address all the points raised by the reviewers. We hope the editor and the reviewers agree that our revised manuscript is satisfied for publication on *Nature Communications*. We would be happy to address any further comments they may have. The revised content mainly contains:

1. Better Clarification of Novelty

We have thoroughly revised our Abstract, Introduction, and Discussion sections to more prominently emphasize the novelty of our work. Moreover, we have included a comprehensive comparative analysis between Att-Transformer and ViT in Methods section and Supplementary. In our responses to the editor and reviewers, we have provided additional detailed explanations regarding our innovative advancements in both hardware and software.

2. Additional Experiments and Details to Justify Our Methodology

To validate our methodological design, we have performed a series of ablation studies to systematically validate the critical role of pretraining and to justify the selection of $k=200$ instances in the first-stage. Furthermore, we have constructed synthetic datasets characterized by varying positive foreground ratios to perform comparative sensitivity analyses of positive foreground ratio on our method and comparative methods. In addition, we have analyzed the influence of the cellular clustering pattern on the stage-one model's classification performance.

3. Additional Experimental Results to Demonstrate Effectiveness of Our Method

We have collected HPV-positive slides ($n=570$) and conducted a sub-study assessing the integrating potential of our approach with HPV screening. We further present our model's performance on detecting low-risk ASC cases. Moreover, we analyze false positives and false negatives in our test set and discuss their clinical implications. To improve model interpretability, we employ class activation maps (CAMs) to visualize the decision-making decision logic.

4. Hardware Optimization and System Validation

Through systematic hardware refinements, we improved the usability of our system, by fully automating slide-scanning process. Engineering schematics of the automated workflow and a demonstration video are integrated in supplementary materials. Subsequently, using multiple statistical metrics combined with interpretation criteria, we have systematically evaluated imaging performance of our compact microscope and the potential impact of image degradation.

5 Additional Files

We have substantially expanded our Supplementary (15 figures and 3 tables now) to support methodological details. Additionally, we have included a video (Supplementary Movie) to illustrate the entire scanning process and internal details of our compact microscope.

Further, to enhance the clarity, we have systematically restructured the Methods section with enhanced modular organization and expanded methodological details. The revised Discussion section now features: (1) A systematic comparison elucidating fundamental discrepancies between cytology and histology, particularly regarding their performance differentials in cervical cytopathology analysis; (2) Comprehensive evaluation of our framework's clinical efficacy in high-risk screening versus its current limitations in ASC detection and cross-scenario generalization; (3) Practical integration pathways with HPV screening protocols. In addition, we have tried our best to make substantial language improvements and reorganized the manuscript to enhance overall readability.

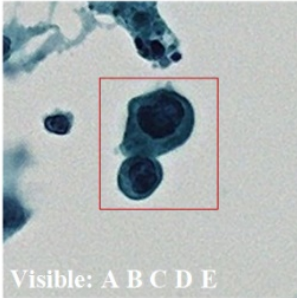
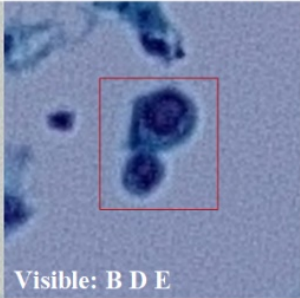
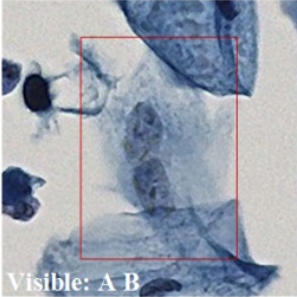
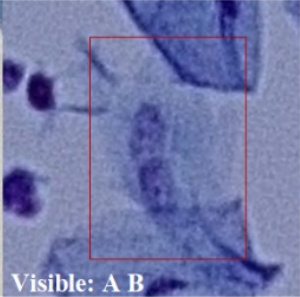
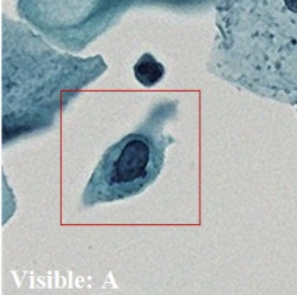
Sincerely yours,

Xiuli Liu (on behalf of all co-authors)

Response to reviewer #1 (Expert in machine learning and digital pathology)

The authors present a cervical cytology precancerous screening system specifically designed for high-risk populations in resource-limited regions. This system integrates a cost-efficient and compact microscope with an AI algorithm to identify HSIL cases. The proposed solution aims to address the bottlenecks in conventional cervical cancer screening, particularly the high costs and reliance on qualified pathologists. The screening system training employs a two-stage slide classification model: 1. The first stage reduces negative cells through self-supervised pretraining, 2. The second stage uses an attention-based transformer (Attention-Transformer) to classify slides by aggregating deep features from instances with the highest lesion probabilities. The system has been validated on a multi-center dataset, demonstrating promising results with high AUC values for screening high-risk SIL cases. This approach has the potential to improve cervical cancer screening coverage in underdeveloped regions, contributing to global health initiatives. Besides the above merits, we have the following points.

1. The development of a low-cost, compact microscope suitable for resource-limited areas, with an approximate pixel resolution of 0.87 $\mu\text{m}/\text{pixel}$ in line 103. I am not sure if the resolution is sufficient to make the diagnosis of other types of cells, for example, LSIL, ASCUS, etc. Or if this is the actual limitation that the system can only use for the high-risk squamous intraepithelial lesion (HSIL).

standard microscope	compact microscope	Annotations
 Visible: A B C D E	 Visible: B D E	HSIL A. The nuclear membrane is irregular. B. The nucleus exhibits hyperchromasia. C. The chromatin is coarse. D. The nucleus is enlarged. E. The nucleus-to-cytoplasm (N/C) ratio is increased.
 Visible: A B	 Visible: A B	LSIL A The nucleus is enlarged. B Binucleation is observed.
 Visible: A	 Visible: A	ASC-US A The nucleus is enlarged.

Response Fig. 1 | Paired images of the same HSIL, LSIL, and ASC-US targets captured using a standard microscope and our portable microscope. The criteria for interpreting these cells are listed on the right. Visible morphological criteria in each image are annotated accordingly.

[Response 1-1] Thanks for the question. The resolution of our portable scanner, 0.87 $\mu\text{m}/\text{pixel}$, naturally prompted reviewers to question the efficiency of our algorithm in recognizing various types of positive cells. To address this, we

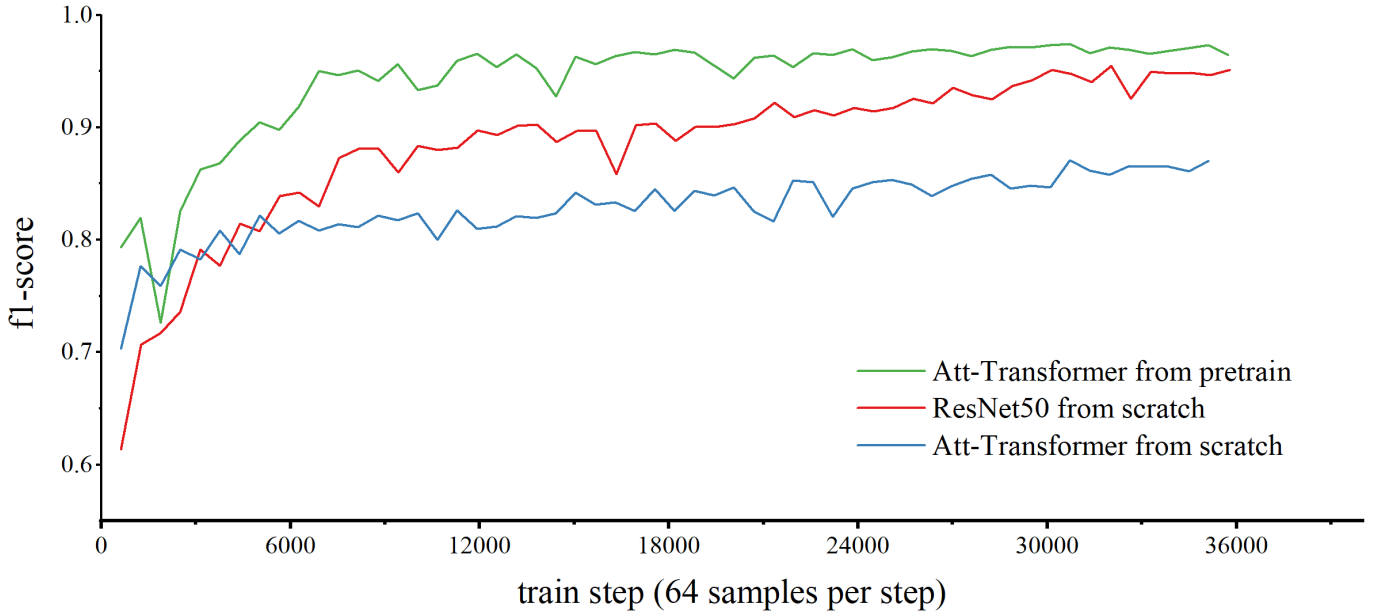
first clarified an ambiguous explanation regarding ‘HSIL’ in the original text. In the manuscript, high-risk cases refer to those categorized as High-grade Squamous Intraepithelial Lesion (HSIL) or Low-grade Squamous Intraepithelial Lesion (LSIL), falling under the Squamous Intraepithelial Lesion (SIL) category according to The Bethesda System (TBS) for reporting cervical cytology. To avoid any misunderstanding, we named SIL rather ‘high-risk’ in the revised version. In experiments distinguishing SIL and negative cases, our method achieved AUCs of 0.889, 0.873 and 0.891 on three external test sets. This result demonstrated its effectiveness in screening for SIL cases. When distinguishing SIL, ASC and negative cases, our method achieved AUCs of 0.744, 0.784 and 0.773 on three external testing sets. The limitations in identifying ASC cases which include ASC-H and ASC-US, mainly attributed to two aspects: the limited resolution of the images and the inherent complexity of interpreting ASC categories.

Compared to standard microscopes, our compact microscope has a lower sampling resolution (0.87 $\mu\text{m}/\text{pixel}$). Response Fig. 1 demonstrates comparative images with the same targets using a standard microscope and our compact microscope. We provide the annotations of the lesions, along with the morphological criteria for interpreting them. Regarding the images of lesions and their morphological criteria, the images obtained with our portable microscope are capable of identifying characteristics such as the size and shape of the cytoplasm and nucleus, as well as the cellular staining intensity. However, its expression is limited in terms of texture and image contrast. The loss of partial morphological features in our images does not hinder the interpretation of holistic cellular pathological changes but may weaken the identification of ambiguous diseases, particularly in the identification of ASC.

It is inherently challenging to clinically distinguish ASC-US. Unlike HSIL and LSIL, which have clear and definite cytomorphological changes, ASC-US typically presents with atypical or subtle cytological morphological features. The loss of details in low-resolution (LR) images makes interpreting ASC even more challenging. What’s more, slides in the ASC category typically contain fewer abnormal cells, which further increases the difficulty of slide-level interpretation.

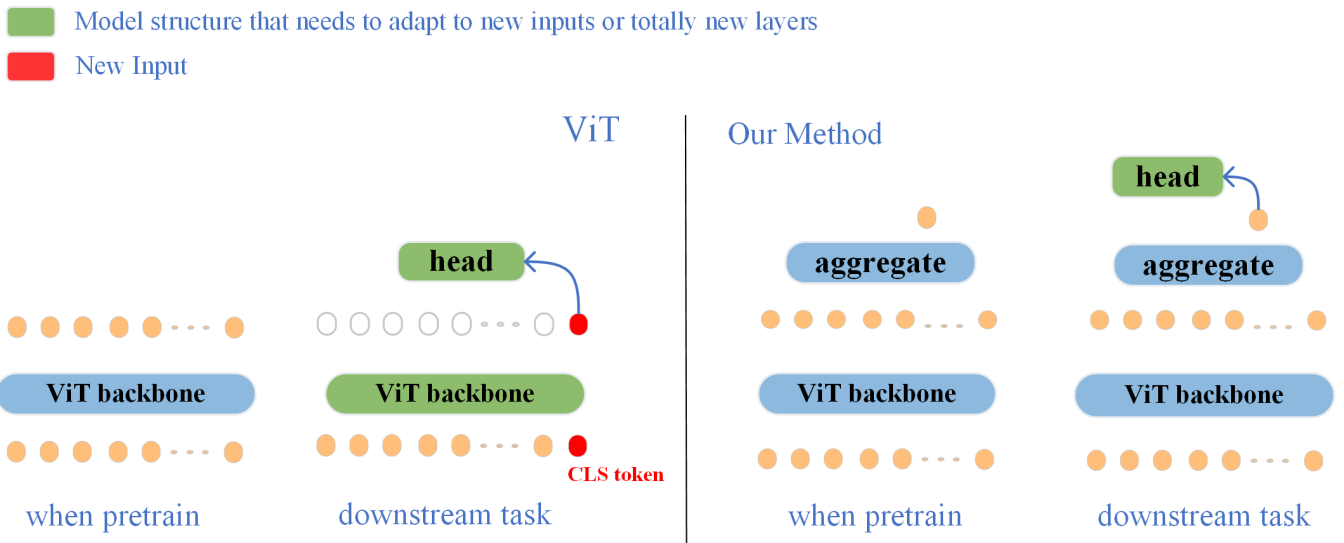
2. The author introduced a dual-stage slide classification model that effectively handles sparsely-distributed lesion cells in cytology whole slide images (WSIs), it is different from the common learning methods, using the supervised for initial training and then apply self-supervised learning to boost the model performance, probably unsupervised learning as last step. Vision Transformers (ViTs) typically demand substantial amounts of training data to achieve competitive performance levels. The original study (Dosovitskiy et al., "An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale," ICLR 2021) concluded that ViTs outperform the top ResNet models across all public datasets, contingent on sufficiently large pre-training datasets. This advantage becomes more pronounced with increasing dataset sizes. When leveraging ViTs with sufficiently large datasets (even for challenging tasks like identifying lesion cells), their performance excels. However, with smaller datasets, ViTs may exhibit slightly lower performance compared to other classification models such as ResNet. The authors need more explain the rationale of the use of self-supervised pretraining on a large number of unlabelled cervical images first, then followed by transfer learning on a smaller set of labeled images.

[Response 1-2] Thanks for the question, the original statement on this point is not clear enough. Indeed, as the reviewer pointed out, using pretrained weights to initialize model parameters is a widely accepted method proven effective across various tasks for enhancing model performance on the target dataset¹⁻³. Compared to convolutional networks, Transformer models typically rely more heavily on proper parameter initialization to achieve optimal results. In our work, we utilize unlabeled instances and the MAE⁴ pretraining method to achieve effective parameter initialization. We demonstrate the effectiveness of our pretraining approach through ablation experiments. As shown in Response Fig. 2, we compare f1-scores on test set across three models: convolutional networks without pretraining (ResNet50 from scratch), our Att-Transformers initialized with pretrained parameters (Att-Transformer from pretrain), and our Att-Transformers without pretraining (Att-Transformer from scratch). All models were trained under same configuration and training dataset, and evaluated using the same test set (test set details: Response Supplementary Data). The results demonstrate that Transformer models initialized with our pretraining method exhibit significantly improved performance, surpassing convolutional networks without pretrained parameters.



Response Fig. 2 | Ablation studies validate our pretraining efficacy. F1-score performance (threshold = 0.5) of our Att-Transformer model with and without pre-trained initialization on the test set throughout training, and that of a classic convolutional network without parameter initialization was also evaluated.

3. The explanation of the Attention Transformer’s specific modifications compared to traditional transformers is brief. More detailed architectural insights would help us understand its unique contributions.



Response Fig. 3 | A schematic diagram of the transfer process of pre-trained parameters to downstream models for both the traditional ViT and our Att-Transformer after pre-training. The green sections represent the model structures that need to adapt to changing inputs in downstream tasks or are randomly initialized. The red sections represent entirely new inputs added in downstream tasks.

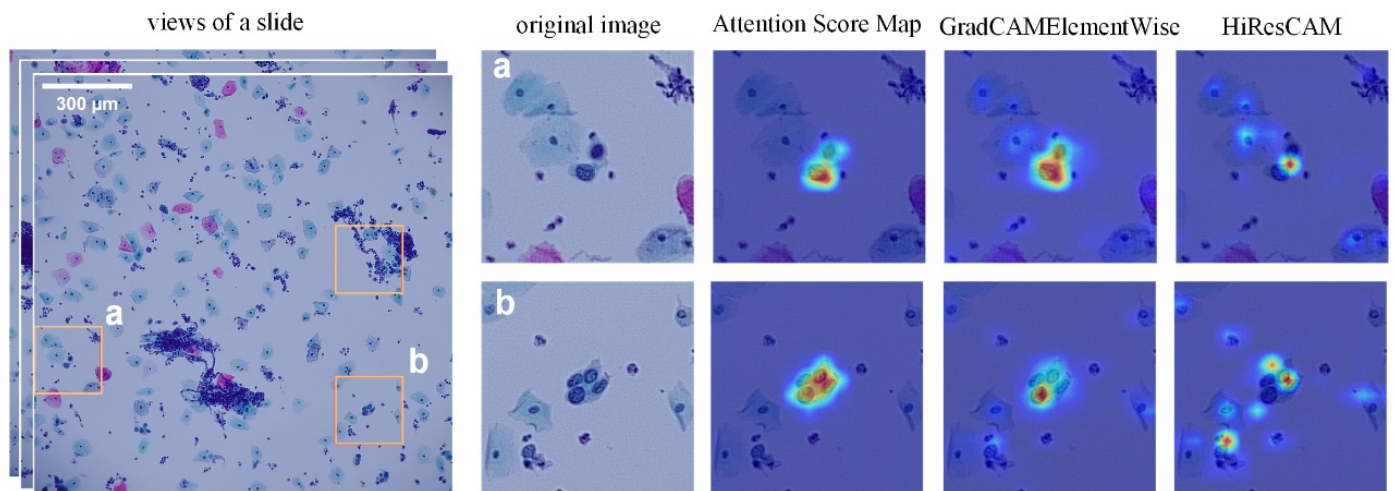
[Response 1-3] We thank the referee for the suggestions. In the revised manuscript, we have added detailed explanations regarding Att-Transformer’s specific modifications, which can be seen in an independent section ‘Att-Transformer’ in Method.

The primary objective of the model design is to emphasize the features of positive lesion areas, enabling the model to make predictions based on these key lesion regions. Compared to traditional transformers, our improvement in the Transformer model lies primarily in how to process the output token sequence after multiple layers of multi-head self-attention as shown in Response Fig. 3. Specifically, we apply a weighted summation to extract information from all

tokens, rather than relying on the CLS token as the final output feature, as in ViT⁵. The weight for each token is generated by a learnable network mapping⁶.

In a transformer architecture, an image is divided into small patches through a sliding window, each mapped to one element in the input sequence. The regions containing positive cells play a decisive role in determining the interpretation result for the entire image. Similarly, only the features of a few instances containing positive cells are decisive for determining the slide's category. Our Att-Transformer effectively focuses on these critical tokens, highlighting their features within the feature sequence. As visualized in Response Fig., high attention scores regions align with lesion-sensitive areas, validated by the attention score heatmaps.

From the perspective of model optimization, in original MAE pretraining, the CLS token of ViT does not participate in pretraining and is introduced only during downstream tasks. Consequently, the pretrained parameters are not aligned with the CLS token, leading to less smooth transfer learning process. In our model, no structural adjustments are needed for the backbone of the model during the transition from pretraining to supervised training, and this design ensures a smooth parameter transfer process.



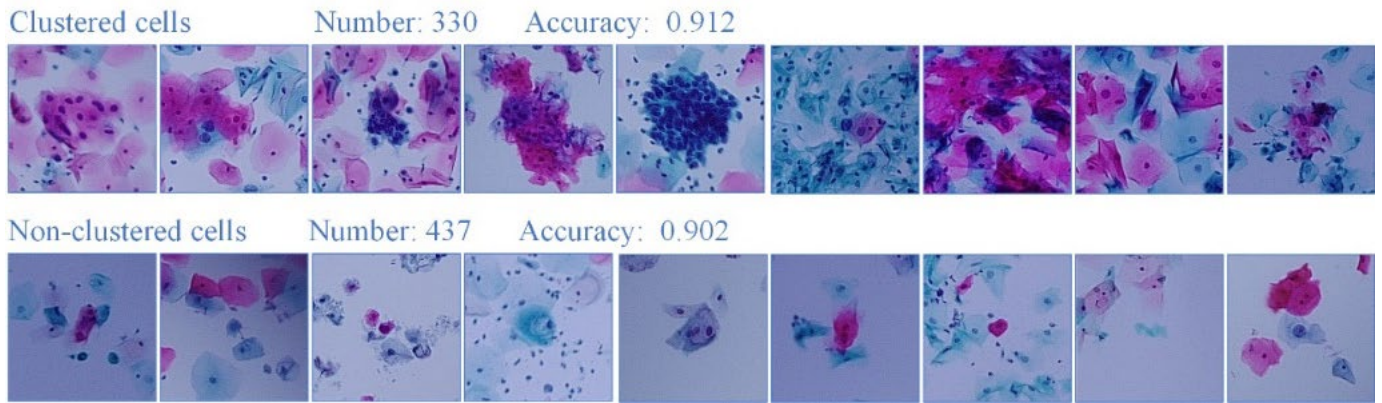
Response Fig. 4 | Visualization of attention mechanism and interpretability. Cell-level ROI and instance-level ROI are acquired from the attention score of the first and second stage models. We frame out instances with the top 200 positive probabilities as instance-level ROI from a view of a slide and plot the heat map of our attention score in stage one. Additionally, class activation maps (CAMs) generated from the output of the penultimate multi-head self-attention layer of our model are plotted, specifically for positive class. Two types of CAM are shown: GradCAMElementWise⁷ and HiResCAM⁸.

4. The authors mentioned the implementation of an attention-based transformer for aggregating deep features from high-probability lesion instances to improve slide classification accuracy, while the feature extract is not well reported in the manuscripts. BTW, we understand that it is often that the cells may form clusters which may bring another challenge to the accurate diagnosis. Is this scenario addressed using the same set of features? Or such clusters were filtered out by the model?

[Response 1-4] Thank you for your insightful questions. The details of the inference process are as follows. The inputs to the first stage model consist of all instances generated from each slide after preprocessing. The outputs of the first-stage model are 768-dimensional features and the positive prediction scores for each input instance. From all the instances of each slide, we select the top 200 instances with high positive prediction scores. The features of these instances, with dimensions of 200×768, are used as the input of the second stage for classifying slides. When test on our low-resolution instance test set, the stage-one model achieved accuracy of 0.99 and 0.91 for negative and positive instances, respectively.

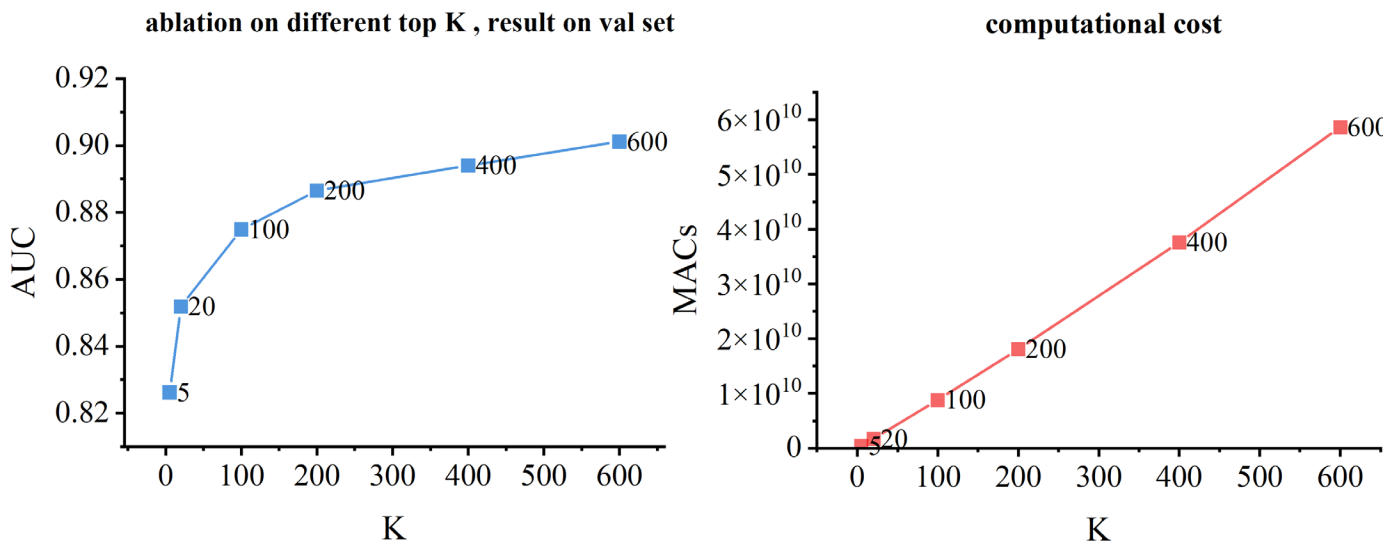
Regarding the feature extraction in stage one, we neither treated instances containing clustered cells as a distinct category, nor excluded them. To evaluate the effect of cell clustering, we divided all positive samples into two groups: clustered cells (330 samples) and non-clustered cells (437 samples). As shown in Response Fig. 5. the model exhibited minimal fluctuation, with accuracies of 0.912 and 0.902 on two subsets. On cervical cytological smears, some cells

typically appear as aggregations or clusters. Clustered cells are not listed as a separate category in our training dataset, but our dataset includes both normal and abnormal cluster cells. Notably, in our classification task, the degree of cell clustering has less impact on the results compared to tasks like instance segmentation or semantic segmentation.



Response Fig. 5 | Two subsets of instances: clustered and non-clustered. The clustered subset contains 330 instances, while the non-clustered subset contains 437 instances. The figure illustrates several instances randomly selected from the two subsets.

5. In line 109, the authors state, "Deep features extracted from redundant instances with the top 200 lesion probabilities are aggregated." Could the authors clarify why the top 200 was chosen specifically? Why not the top 100, top 300, or another number? Considering there are tens of thousands of cells in the whole slide images (WSI) of cervical cytology as mentioned in line 79, the rationale behind selecting this specific number would be valuable for understanding the methodology. The paper does not sufficiently detail the criteria for selecting the top 200 instances with the highest lesion probabilities. Clarification on this selection process would enhance reproducibility.



Response Fig. 6 | Instance features of different top k were used as the features of slides to train the slide classification model. The figure shows its area under the receiver operating characteristic curve (AUC) on the validation set. Meanwhile, we presented the computational cost of model inference when different K values were used, which was estimated by the number of multiply-accumulate operations (MACs).

[Response 1-5] Thanks for the suggestion. Our statistical results indicate that the proportion of positive cells reaches a maximum of 0.01, and is typically on the order of 0.001. Therefore, if an ideal first stage classifier were trained, it could select a very small number of top-ranked instances, aligning with the design logic of many traditional algorithms⁹⁻¹¹. Cheng et al. used top 10 instances, for example⁹. In cases where it is not feasible to train an ideal first stage classifier, we opt to extract features from the instances with the high probabilities of potentially being positive, and in the second

stage, the model further classifies slides based on these extracted features. We first excluded the approach of using features of all instances as the slide features. Such features exhibit poor specificity between positive and negative slides due to the low proportion of positive instances. By selecting 200 instances, we ensure coverage of all potential positive features. This approach offers two advantages. First, we reduce the number of negative instances in the first stage, thereby enhancing the specificity of the slide feature. Second, the second-stage model can leverage the slide labels to learn the complex decision process that maps the features of potential positive instances to the overall slide interpretation.

We designed experiments to investigate the effect of using different top-k instances on the performance of the slide classification model, as shown in Response Fig. 6. The results demonstrate that models using a larger k achieve superior performance. However, increasing the number of exhibits poor specificity instances also exhibits diminishing marginal returns, particularly when considering the computational cost. We calculated the inference computational cost of models with varying number of instances, and within the range of 600, the computational cost of inference shows an approximately linear correlation. In summary, selecting k=200 represents a trade-off.

* Note on computational scaling: The length of the input sequence further increases, the computational cost will gradually follow a square growth pattern. The backbone of the Att-Transformer is composed of stacked Multi-Head Self-Attention (MHSA) layers. The cost for a single MHSA layer (MACs) is:

$$\text{COST} = L \times (4 \times F \times D) + L \times (2 \times L_0 \times D) + L^2 \times (2 \times F)$$

where L is the length of input sequence, F is the dimension of q, k and v in transformer, D is the dimension of input token, L_0 is the number of neurons in the intermediate layer of MLP. As seen from the formula, the computation cost includes two linear terms (L) and one quadratic term (L^2). For smaller L , the linear terms dominate, for larger L , quadratic term scaling prevails.

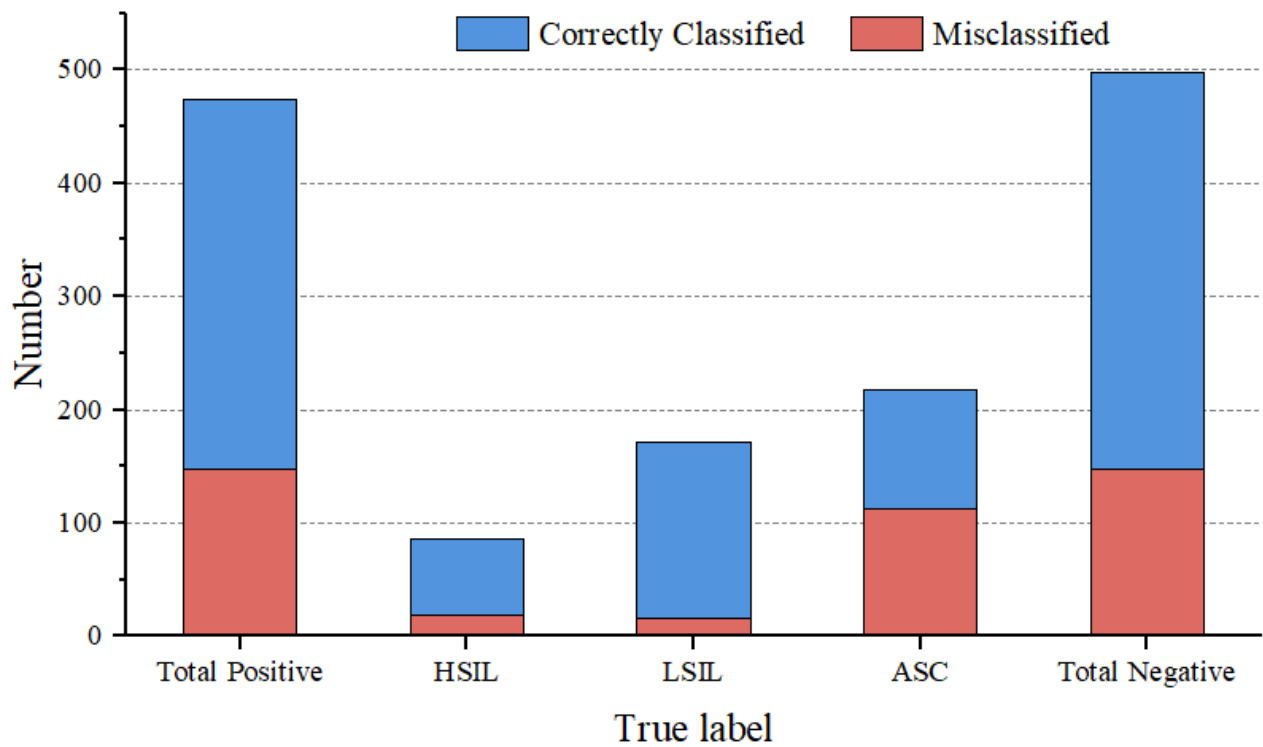
6. While the final performance metrics are provided, there is a lack of discussion on potential false positives and false negatives and their implications on clinical decisions. In addition, some important intermedia results will also be appreciated if included in supplementary material.

[Response 1-6] Thanks for this suggestion. To analyze the impact of potential false positive and false negative cells, we present model's classification results for each subclass in Response Fig. 7. We analyze the prediction results from three external testing sets and find that most false negative slides were misclassified ASC slides. Our method demonstrates the lowest misclassification rate for LSIL, followed by HSIL, and the highest misclassification rate for ASC.

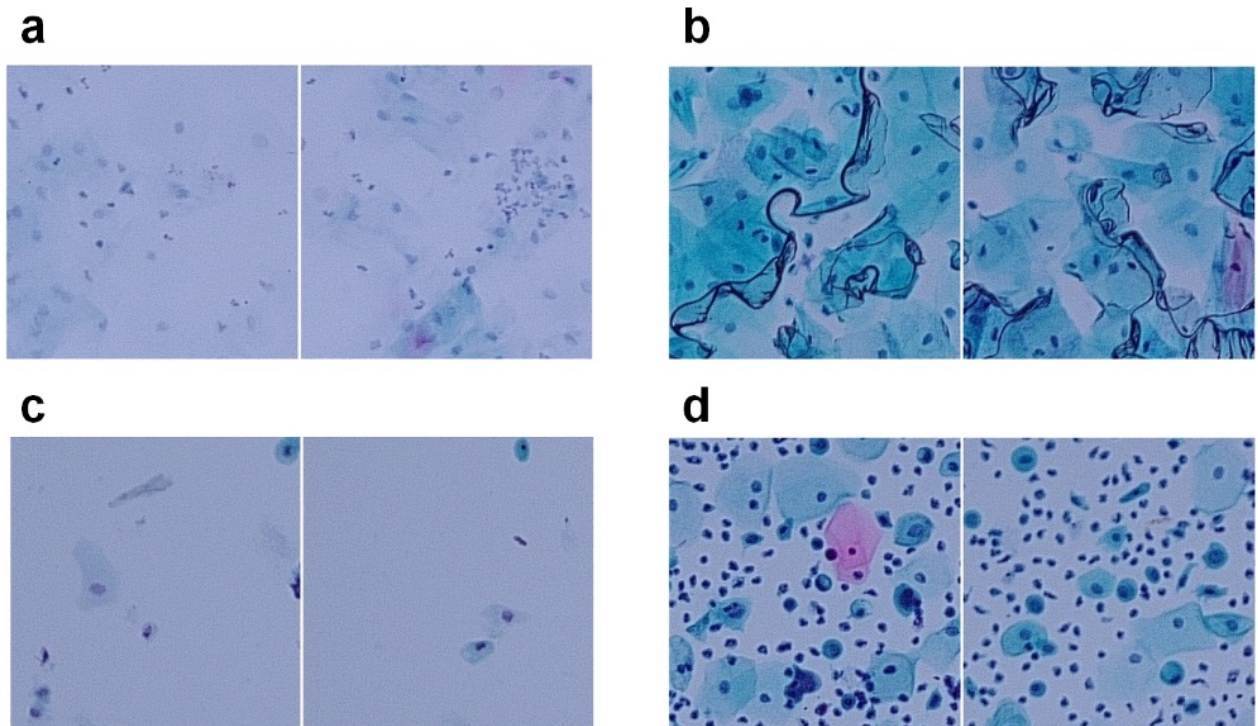
Clinically distinguishing ASC-US is inherently challenging due to its atypical or subtle cytomorphological features, unlike the clear changes seen in HSIL and LSIL. Low-resolution (LR) images exacerbate this challenge by losing critical details, and the scarcity of abnormal cells in ASC slides further complicates slide-level interpretation. In clinical practice, SIL (HSIL and LSIL) is the most important types of abnormalities to identify, as it shows clear cytomorphological changes that suggest a higher risk of cervical cancer. ASC, in comparison to SIL categories, represents lesions that are either atypical or insufficiently pronounced. In clinical practice, ASC categories also exhibit significant inter-observer variability, with a relatively lower risk of malignancy. Further examination or follow-up is typically recommended for ASC. For screening in resource-limited settings, our method is effective in identifying SIL, which is the most important screening indicator. However, its lower ability to detect ASC lesions means that our method may not be suitable for the detection of earlier-stage cancers.

Additionally, several other factors may influence the performance of the model, as illustrated in Response Fig. 8. The slide quality in primary hospitals is generally inferior to that in central hospitals, potentially exhibiting issues such as inadequate staining, insufficient cell count, or the presence of artifacts like bubbles. Moreover, inflammatory reactions can significantly alter the image style of the slides, potentially reducing the model's interpretative accuracy. These are issues that warrant further investigation in future studies.

In the context of cervical cancer screening, false positives may lead to unnecessary further testing for patients. In contrast to false negatives, this type of misjudgment does not carry the risk of missing diagnoses.



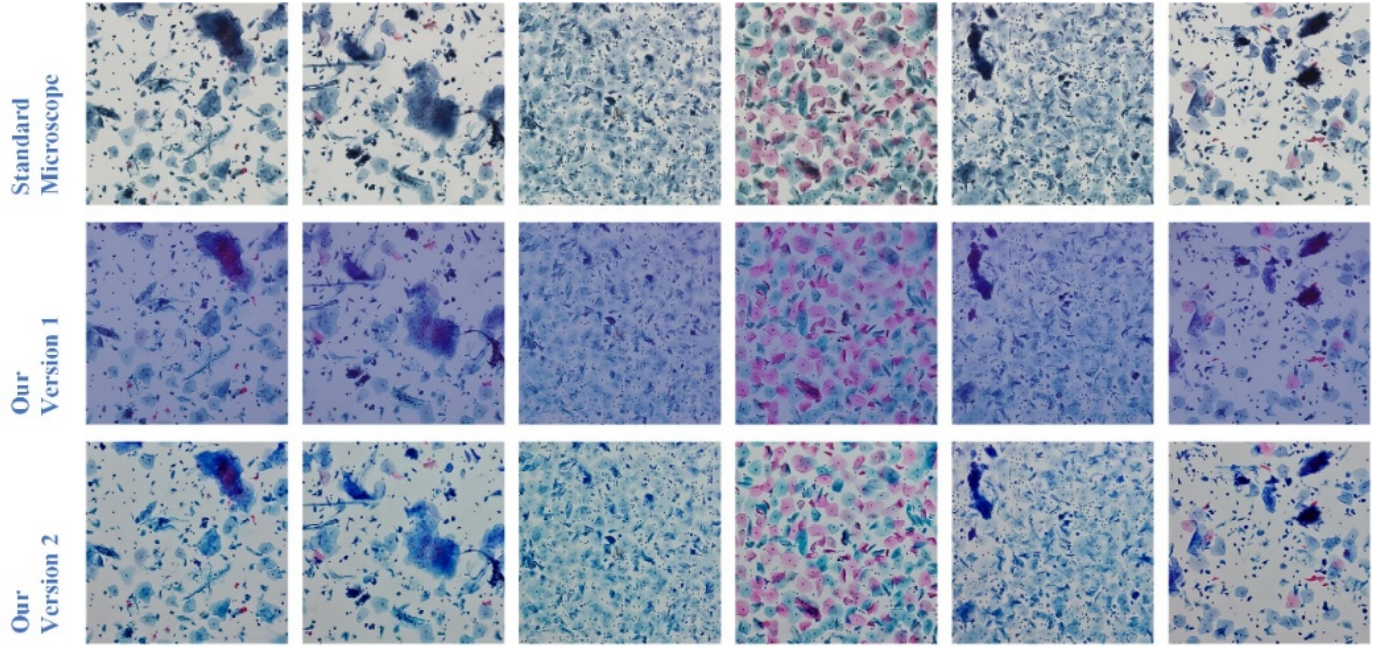
Response Fig. 7 | The model's classification results for each subclass (threshold = 0.5), using data from three external test sets.



Response Fig. 8 | Image quality issues found in some incorrectly interpreted slides. a) Under-staining or fading b) Bubble artifacts c) Insufficient cell count d) Inflammatory reaction.

7. Authors acknowledge the variability in slide preparation and staining processes across different regions, which can affect AI algorithm generalization. The visual comparison was provided in Figure 1b and a supp figure, while a quantitative comparison will be better to understand the difference between the images and how they impact the diagnosis. For example, the authors mention providing 100 pairs of images for an imaging quality check in line 153. It would be beneficial to include quantitative measurements, such as the percentage of lost focus, the frequency of

saturation, noise levels, etc., for a more comprehensive evaluation of the images generated by the compact microscope. Currently, a detailed assessment of the generated images is not provided.



Response Fig. 9 | Imaging results of the same target using a standard microscope, our first version microscope, and our second version microscope.

Response table 1. metrics for image quality			
metric	Standard microscope (HR)	Our version 1 (v1)	Our version 2 (v2)
PSNR (with HR)	\	15.76	19.89
SSIM (with HR)	\	0.497	0.644
Noise Std Dev	5.49	19.63	4.24
percentage of lost focus ¹²	0.013	0.017	0.015
Saturation Probability	0.0374%	0.0000%	0.0115%
Resolution (1951 USAF)	912.3 lp/mm (0.54μm)	287.4 lp/mm (1.74 μm)	574.7lp/mm (0.87 μm)

[Response 1-7] Thank you for the valuable comments. We have provided visual comparisons in Figure 1b as an illustrative example. We included Response Fig. 9 and Response table 1 to further quantify these differences.

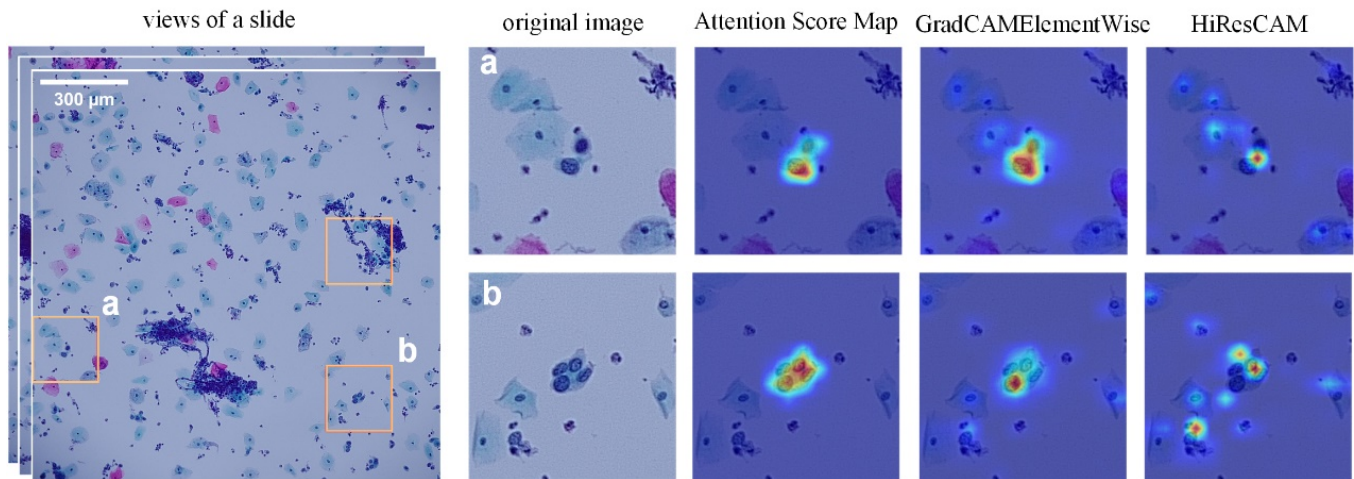
Three microscope configurations were used to image the same object: a standard microscope, the first and the second versions of our designed microscope. A total of 42 image pairs (pixel dimensions: 3000×3000) were acquired for image quality assessment, with representative samples displayed in Response Fig. 9.

We acknowledge that a systematic comparison are critical for evaluating imaging differences between images and their diagnostic implications. As shown in Response table 1, First, we quantified key parameters including the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) values across both versions of our microscope and the standard microscope. To calculate areas of lost focus, we used the defocus estimation map¹² from Chen et al. and applied a threshold segmentation method (threshold = 0.7) to obtain the area ratio of defocus in the image. Furthermore, we calculated the probability of saturated pixels appearing in the images. To evaluate noise levels, we selected three 100×100 pixel white background images and computed the standard deviation of pixel intensity for the three imaging modes. Moreover, we captured images of 1951 USAF resolution calculator using the two versions of our designed microscope to determine the actual resolution of the images.

The results indicate that the second iteration of the microscope achieves higher PSNR and SSIM values compared to the first iteration. However, both custom-designed microscopes exhibit substantial gaps in texture fidelity relatively to the standard microscope. Despite the apparent lower brightness of the first version's images, the probability of saturated pixels is significantly low across all three imaging modes. Notably, the optical resolution of the second version of the microscope is $0.87\text{ }\mu\text{m}$, which shows a significant improvement over the first version with $1.74\text{ }\mu\text{m}$. Overall, both versions of the microscopes show no significant difference from the standard microscope in terms of defocus area and saturated pixel ratio. The noise level of the first version of the small microscope is relatively higher, while the noise level of the second version is comparable to that of the standard microscope. The optical resolution of both versions of the small microscopes is inferior to that of the standard microscope, with the second version showing improvement over the first.

* Clarification: In the initially submitted manuscript, the resolution of the first version of the compact microscope was $2.19\text{ }\mu\text{m}$, that is different from the current value of $1.74\text{ }\mu\text{m}$. This discrepancy arose from limitations in the insufficient range of our original 1951 USAF resolution calculator.

8 Although the Attention Transformer provides some level of interpretability through attention scores, further efforts to enhance model transparency and explainability would be beneficial. Techniques such as saliency maps or other visualization tools could be employed to provide clearer insights into the model's decision-making process.

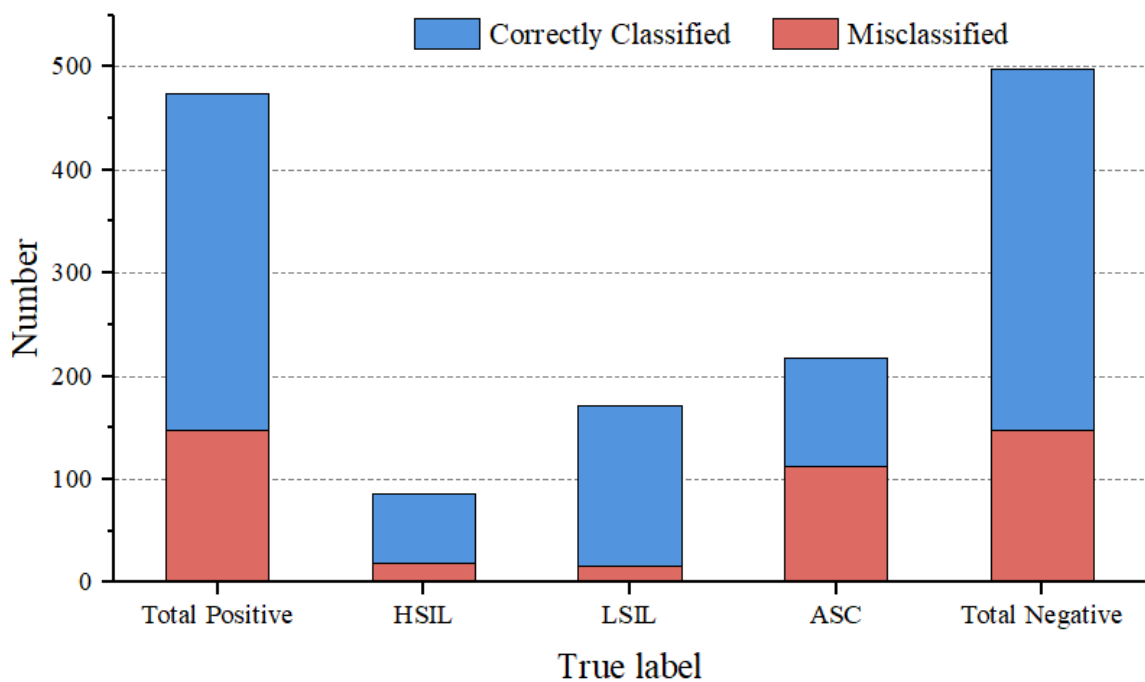


Response Fig. 10 | Visualization of attention mechanism and interpretability. Cell-level ROI and instance-level ROI are acquired from the attention score of the first and second stage models. We frame out instances with the top 200 positive probabilities as instance-level ROI from a view of a slide and plot the heat map of our attention score in stage one. Additionally, class activation maps (CAMs) generated from the output of the penultimate multi-head self-attention layer of our model are plotted, specifically for positive class. Two types of CAM are shown: GradCAMElementWise⁷ and HiResCAM⁸.

[Response 1-8] Thanks for the suggestion. We agree that while the Attention Transformer provides some level of interpretability through attention scores, further enhancing the model's transparency and explainability would be very beneficial. As shown in Response Fig. 10, to better understand the model's decision-making process, we make use of the attention scores output by our model to generate Attention Score Map. We also generated CAMs based on the feature maps from the penultimate multi-head self-attention (MHSA) layer. We choose GradCAMElementWise⁷ and HiResCAM⁸ to plot the map. All three heatmaps consistently focus on the lesion regions within the images and the results demonstrate high degree of similarity between the two CAMs and the attention map generated using our attention scores. The heatmaps generated by our method achieve the best performance in identifying sensitive lesion regions. In Response Fig. 10, the lesion regions primarily consist of cells with abnormally enlarged nuclei.

In the feature sequence output by the model backbone, our attention scores assign higher weights to the tokens corresponding to the lesion areas, allowing the final output features of the model to be primarily determined by the lesion regions. Meanwhile, the ROIs identified by the heatmaps closely align with the ROIs identified from a pathological perspective, indicating that the lesion areas emphasized by our method contain reliable semantic information.

9 The study successfully uses low-resolution WSIs, but there are inherent limitations in detecting fine details, such as low-grade lesions or atypical cells, which is sometimes more important for early detection. Not sure what is the performance of those cell types.



Response Fig. 11 | The model's classification results for each subclass (threshold = 0.5), using data from three external test sets.

[Response 1-9] Thank you for your suggestions. Indeed, while our study successfully utilizes low-resolution whole slide images (WSIs), there are inherent limitations in detecting fine details. We analyze the prediction results from three external testing sets and find that most false-negative slides were misclassified ASC slides as shown in Response Fig. 11. Our method demonstrates the lowest misclassification rate for LSIL, followed by HSIL, and the highest misclassification rate for ASC.

In experiments distinguishing SIL categories (HSIL and LSIL) from the negative category, our method achieved AUCs of 0.889, 0.873, and 0.891 across the three test sets as shown in Figure 4 (Figure 3 in revised version) in the manuscript. When including the ASC category, our method achieved AUCs of 0.784, 0.773, and 0.744 respectively. These outcomes demonstrate the effectiveness of our approach in screening for SIL types. In clinical practice, SIL categories are considered to have a higher risk of malignancy due to clear cytomorphological changes. In contrast, the ASC category presents a relatively lower malignancy risk, characterized by atypical or insufficient cytomorphological changes. Patients with ASC are typically advised to undergo follow-up. Our current research focuses on screening high-risk populations of cervical cancer (SIL cases) through our compact microscope and algorithm in resource-constrained areas.

10. The manuscript would benefit from the expertise of a professional editor to enhance the language. Currently, it is challenging to read, piecewise lacks detailed information, and is not well organized. Additionally, the supplementary documentation does not adequately aid in understanding the experiment and implementation.

[Response 1-10] We are fully aware of the deficiencies in writing the manuscript. We are deeply grateful for the valuable feedback provided by the reviewers. In response, we reorganized and expanded the content of the Methods section, and thoroughly revised errors and unclear descriptions throughout the manuscript. Additionally, we have conducted multiple sets of ablation experiments and incorporated several statistical metrics to further refine our work.

We first reorganized our abstract to emphasize the innovativeness of our approach. In the introduction, we added comments related to HPV nucleic acid testing and clarified the relationship between histopathology and cytology pathology.

In the results section, we conducted a more comprehensive and systematic evaluation about the image quality of our microscope based on statistical metrics and visualization plots. Then, we further explained the innovativeness of our method compared to ViT, as well as the advantages of our approach. We added ablation experiments to explain the rationale behind selecting 200 positive example features as slide features in the first stage. We also discussed the model's performance in the ASC category and its clinical significance. We conducted a sub-study on HPV-positive slides. Additionally, we further elaborated through experiments on the significant impact of the positive foreground ratio on the slide classification task.

In the discussion, we discussed the issues of false positives and false negatives encountered during the experiments, and supplemented the potential application of our method in conjunction with HPV nucleic acid testing for screening.

We reorganized the method section to make it logically clearer. We used formulas to elaborate on the details of the Att-Transformer we proposed. We separately described the pretraining principles and process of the first-stage model. Then, we described the training processes for the first and second-stage models. We also explained the principles behind identifying the ROI using our method. Finally, we revisited and clarified all necessary information regarding the datasets used in the experiments.

We significantly expanded our supplementary material. The new supplementary, with 15 figures and 3 tables, is used to illustrate the various points mentioned above. The supplementary also includes various metrics of the improved system.

11. Some recent publications on this topic as follows are not cited, it is also important to differentiate this work from the existing works:

[Response 1-11] We thank the referee for this critic. We have reviewed recent papers, and find that the two representative papers you proposed are highly related to our work, which are cited as shown below:

“35. Wang, C.-W. et al. Artificial intelligence-assisted fast screening cervical high grade squamous intraepithelial lesion and squamous cell carcinoma diagnosis and treatment planning. *Sci Rep* 11, 16244 (2021).”

“39. Zhu, X. et al. Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears. *Nat Commun* 12, 3541 (2021).”

“40. Wang, J. et al. Artificial intelligence enables precision diagnosis of cervical cytology grades and cervical cancer. *Nat Commun* 15, 4369 (2024).”

“41. Zhang, Z. et al. SCAC: A Semi-Supervised Learning Approach for Cervical Abnormal Cell Detection. *IEEE Journal of Biomedical and Health Informatics* 28, 3501–3512 (2024).”

Minor points: reviewer #1

1. In line 184, “The right figure shows the average pixel intensity”, it is clear that the y-axis represents pixel grey intensity. What is the x-axis? Micro or pixel?

Re: Thanks, we have corrected as suggested. The x-axis is by pixel.

2. In line 196, citation number 40 cannot be found in the manuscript, as only 37 citations are listed. Please check all citations and references.

Re: Thanks, we have checked all citations and references and have corrected as suggested.

3. Could you please clarify the difference between cell level and instance level in Figure 2? Are they referring to the same concept? Please consider adding the dimensions of the patches (224x224 and 256x256) in the figure for clarity. Ensuring that the main text and the figure complement each other will aid the reader's understanding. Currently, it is unclear which is the first and which is the second stage in the figure. While the figure is presented, it is not helpful.

Re: Thanks, we have modified Figure 2 to make it clear.

4. In line 483, Camelyon16 reported a dataset comprising 400 WSIs. The first training dataset includes 170 WSIs of lymph nodes (100 normal and 70 containing metastases), while the second consists of 100 WSIs (60 normal and 40 containing metastases). The test dataset comprises 130 WSIs collected from both universities. However, there is uncertainty regarding the standard histology of 110 WSIs and the criteria for their selection.

Re: Thanks. To calculate the proportion of positive foreground, we only used slides containing positive foreground from the training set, a total of 110 slides.

5. In line 484, The process of creating 106 fully annotated standard WSIs is unclear. Specifically, details such as the number of cytologists involved, and any cross-checks and validation procedures for the annotated data, are not specified.

Re: Our fully annotated dataset includes comprehensive annotations of positive cells present in each slide, completed independently by three experts. Then, the three pathologists discuss the discrepancies and reach a consensus.

6. In line 491, the process from 224 pixels to obtain 16 pixels in this study is unclear; the sentence appears to be incomplete.

Re: Thanks, we have reorganized relevant content.

7. Could you please provide details regarding the accuracy in detecting negative and challenging samples described in line 522(552)? Additionally, could you elaborate on the validation process used to identify these challenging samples?

Re: Thank you for your question. We show more details about our false positive mining training (FPMT) at Finetune for stage one in Method.

8. In line 561, upon reviewing the references cited in this sentence, I couldn't locate any related sources, including supplementary materials, regarding this augmentation process. I kindly suggest considering the addition of more detailed information to this manuscript.

Re: Thank you for your feedback. We have made a careful revision to ensure to include more comprehensive sources and additional details regarding the augmentation process in Finetune for stage one in Method. These augmentations primarily include color jittering in the RGB and HSV color spaces, Gaussian blur, contrast enhancement, as well as rotation and symmetry operations.

9. Improving the formatting of subsections would be beneficial, as exemplified in lines 463, 480, 527, 536, and so forth.

Re: Thanks, we have corrected as suggested.

The paper presents a promising approach to cervical cytology precancerous screening in resource-limited regions, leveraging a low-cost compact microscope and an AI-based classification system. While the results are encouraging, some above issues need to be well addressed.

Re: Thank you for your positive feedback and constructive comments. We acknowledge the issues you have raised and are committed to addressing them thoroughly. We have tried to revise the manuscript accordingly, ensuring that all concerns are adequately addressed.

Response to reviewer #2: (Early-Career Researcher co-reviewer)

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Response to reviewer #3: (Expert in microscopy, imaging, and machine learning)

This paper developed a cervical precancerous assisted-screening system for identifying high-risk squamous intraepithelial lesion cases, which was composed of a low-cost miniature microscope and a low-pathologist-reliance artificial intelligence algorithm. Though the research is interesting and may have a wide clinical application, the novelty and the presentation of this work is still limited. In addition, I have following other concerns:

1. The abstract part fails to reflect the novelty of the paper. Please further describe the hardware and software innovations of the developed system in detail.

[Response 3-1] We thank the referee for pointing out this issue. This paper presents a cervical precancerous assisted screening system for identifying high-risk populations. The system's innovation lies in both its hardware and software components, detailed as follows:

Traditional screening methods rely on pathologists to interpret glass slides using high-performance imaging devices, with or without AI-assisted tools. These screening methods have achieved great results in developed countries and regions with adequate medical resources, significantly reducing the incidence of cervical cancer in these areas. However, conducting screenings in underdeveloped regions remains a bottleneck in global cervical cancer screening, as the incidence rate in these areas remains high. The low coverage of cervical screening in underdeveloped regions can be attributed to two main factors. On the one hand, there is a lack of supporting medical facilities, which are costly, sophisticated, and require long-term maintenance. On the other hand, there is a shortage of experienced pathology experts in these areas. We propose a feasible solution, which includes a low-cost and compact imaging hardware, along with AI algorithms adapted to the imaging quality of such low-cost devices, ultimately enabling effective screening for high-risk populations of cervical cancer.

From the hardware perspective, our portable scanner utilizes an architecture based on consumer level electronics. Compared to high-performance laboratory equipment, it allows for easy procurement of components while reducing the overall cost by tens to hundreds of times. The total cost of our hardware is kept under \$300. For details, we utilize an optical path design based on plastic aspheric mirrors, which not only reduces the optical path length but also significantly compresses the device's overall size. These designs ensure that our system is highly accessible and portable while maintaining imaging quality sufficient for identifying SIL (squamous intraepithelial lesion) categories of pathology. These hardware features provide an excellent foundation for conducting screening activities in resource-constrained settings.

The current AI-assisted cervical cancer screening algorithms detect a small number of positive cells, ranging from several thousand to ten thousand with high accuracy, because there are typically only a very small number of lesion cells in positive slides. For example, Cheng et al. used ten suspicious cells to predict the slide label¹³. In these methods, the efficacy of slide classification is highly contingent upon the precise identification of abnormal cells, which necessitates the utilization of high-resolution images replete with sufficient texture details and large-scale annotation datasets. To leverage the limited imaging quality of our compact microscope, we first exclude a large number of negative examples from the slides and extract features from 200 potential positive examples, thus obtaining a good feature representation of the slide. In the second stage, we then use the slide labels and the extracted slide features to train the final classifier. We introduce the Att-Transformer, which improves the vision transformer by incorporating a learnable attention module to aggregate the feature of each patch in an image, allowing the model to focus on the sparse positive cells in the image that play a decisive role in classification.

We have rewritten the abstract to better emphasize the innovations of our approach in both hardware and software, as shown below:

“Insufficient coverage of cervical cytology screening in underdeveloped and resource-limited areas remains a major bottleneck, as traditional centralized methods require significant investment and many qualified pathologists. By leveraging consumer-grade electronic hardware and non-curved mirrors, we designed an extremely low-cost and compact microscope. Due to the microscope's low resolution, which hinders accurate identification of lesion cells, we

trained a coarse instance classifier to screen and extract feature sequence of the top 200 instances containing potential lesions from a slide. We further developed Att-Transformer to focus and integrate the sparse lesion information from these sequences, enabling slide grading. Our model was trained and validated using 3,510 low-resolution slides from four hospitals, with subsequently evaluation on four independent datasets. The system achieved area under the receiver operating characteristic curve values of 0.87 and 0.89 for detecting squamous intraepithelial lesions on 364 slides from two external primary hospitals, 0.89 on 391 newly collected slides from the original four hospitals, and 0.85 on 570 human papillomavirus positive slides. Overall, our approach demonstrates the feasibility of effectively detecting high-risk cases in resource limited area.”

2. While AI-assisted cancer screening has been successfully applied in several related fields, the authors need to show at least killer applications for the demonstration of their method with demonstration of their unique advantages. Currently, the demonstrations can be achieved by many previous methods.

[Response 3-2] We thank the referee for this critical feedback. Cervical cancer remains one of the major threats to women's health, especially in resource-limited areas. Traditional screening methods rely on pathologists to analyze slides using high-performance imaging devices, sometimes with the assistance of AI tools. These methods have proven highly effective in developed countries and regions with sufficient medical resources, leading to a significant reduction in cervical cancer incidence in those areas. However, screening in underdeveloped regions remains a challenge, as the incidence rate of cervical cancer continues to be high. In resource-limited countries and regions, about 85% of new cervical cancer cases are diagnosed. These areas often face multiple challenges, including insufficient vaccine supply, scarce medical facilities, which make it difficult to effectively carry out screening, prevention and treatment of cervical cancer. For details, the limited coverage of cervical cancer screening in these areas is primarily due to two factors: the high cost and complexity of medical equipment that requires ongoing maintenance, and the shortage of experienced pathologists.

To address these challenges, we propose a practical solution that includes low-cost, compact imaging hardware paired with AI algorithms tailored to the imaging quality of such affordable devices, thereby enabling effective screening for high-risk populations in underdeveloped regions. Due to the low hardware cost, easy availability, and ease of maintenance, the hardware is well-suited for screening activities in remote areas. The AI algorithm effectively performs recommendations for high-risk cervical cancer populations, which reduces the reliance on pathology experts. Our algorithm adapts to images with limited details and successfully identifies SIL (HSIL and LSIL) cases. Through our system, screening can be implemented in resource-limited areas. High-risk individuals can be recommended to advanced screening centers. Additionally, our system can also work in conjunction with HPV nucleic acid testing for joint screening, enhancing the overall specificity and sensitivity (See the response 6-3).

3. Please add what is the unit of the horizontal coordinate on the right picture of Figure 1b.

[Response 3-3] The horizontal axis represents the relative pixel position along the sample line in the left image. To better convey this idea, we have redrawn the Figure 1b.

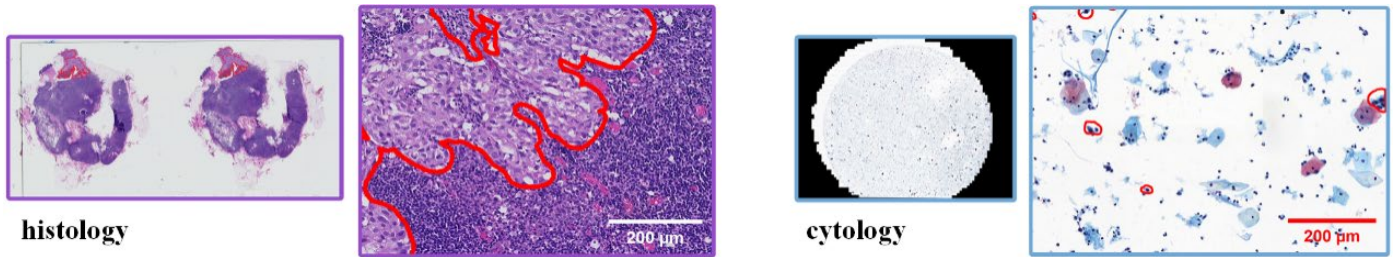
4. It seems that there is a contradiction between the sentence “In computational pathology field, there have been many high-performance weakly-supervised WSI classification methods which are built without the need for cell-level annotations. However, there are few such kind of methods for cytology WSIs” and “On the contrary, most of them need huge number of cell-level annotations as mentioned in the introduction”. Are cell-level annotations needed or not for the mentioned methods?

[Response 3-4] We really appreciate the reviewer’s comments and recognize that our original statement was unclear. As suggested, we have rewritten the corresponding content and would like to clarify our intended meaning.

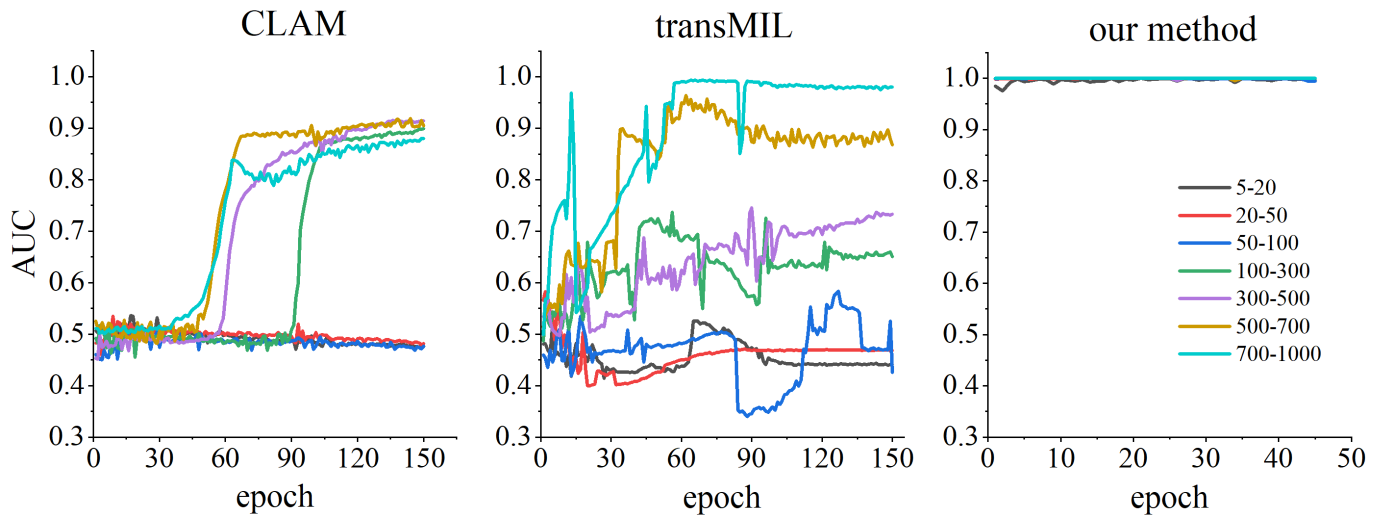
Computational pathology can be broadly divided into two areas based on the objects of study: histopathology and cytopathology. These two areas differ significantly in terms of the characteristics of their study objects and the

mainstream research methodologies applied. For histopathology, the current mainstream approach is based on weakly supervised methods, which rely solely on annotations at the whole-slide image (WSI) level without requiring detailed annotations for specific regions. In contrast, cytopathology primarily employs methods for detection, classification of localized lesion area, or a combination of both, which generally require a large number of annotations at the cellular level, representing a much finer level of granularity than histopathology.

5. The paper repeatedly mentioned that the poor performance of the comparison methods was due to the "limited resolution of our compact microscope" and "sparsely-distribution lesion cells". Please add the corresponding results in the text to show the influence of the above two reasons on the comparison method.



Response Fig. 12 | Examples of a histopathology slide from Camelyon16 and a fully annotated cytopathology slide from our dataset are presented. The area outlined by the red curve indicates the positive foreground.



Response Fig. 13 | We constructed a virtual dataset with varying positive foreground ratios using annotated positive samples and negative slides, and tested the performance of CLAM, TransMIL, and our proposed method on this dataset. The curves in the figure represent the model's performance on the test set during training process.

[Response 3-5] Thanks for the suggestion. To address the reviewer's concerns about the influence of the "limited resolution of our compact microscope" and "sparsely-distributed lesion cells" on the performance of comparison methods, we further elucidate this issue through qualitative reasoning and ablation experiments.

Influence of Sparsely-Distributed Lesion Cells: In whole-slide classification algorithms for histopathology, pretrained models are first used to extract and reduce the dimensionality of instance-level features^{14,15}, and the sequence of features from all instances is aggregated to form a slide-level feature representation for slide classification. In these methods, the proportion of positive instance features in the sequence affects the distinction between positive and negative slide feature sequences, which in turn impacts classification difficulty. Generally, the higher the proportion of positive features in the sequence, the stronger the specificity of positive slide features relative to negative slides. We already know that the proportion of positive foreground in cytopathology data is significantly lower than that in histopathology data. In histopathological data, the lesion areas are large and concentrated, whereas in cytopathological data, the lesion areas are small and dispersed as shown in Response Fig. 12. In our statistical results, the proportion of

positive foreground to all foreground is on the order of 0.01 at its highest, and more commonly on the order of 0.001. The sparsity of positive cells in cytopathology makes it challenging to train models using traditional methods in histopathology.

Influence of limited resolution of our compact microscope: To address the former issue, current algorithms for cervical cytopathology data typically rely on large-scale instance-level annotations to train an accurate instance-level classifiers or detection models for identifying positive instances across the entire slide. This approach requires high-performing models to accurately identify these very sparse positive cells, which puts forward new demands for image quality and datasets with high-quality annotations. It is difficult to train instance-level models with such high performance using the same paradigm based on images with limited image quality captured by our portable microscope.

Based on this, we designed a two-stage method to address these challenges. In the first stage, the model eliminates a large number of negative instances, akin to how pathologists quickly exclude negative regions under low magnification. The first-stage model extracts potential positive instance features and filters negative instance features, resulting in a relatively shorter feature sequence with higher positive proportion that enhances the specificity between positive and negative sequences. We select the top 200 instances, which are sufficient to contain all potential positive cells. Based on this, the second-stage model further uses slide-level labels to learn the mapping between these features and the corresponding labels.

Ablation experiment: We trained and tested the model's classification performance using virtual datasets with varying positive foreground ratios, constructed from positive instances and negative whole slide images. After preprocessing, a negative slide is converted into a large number of negative instances, with an average of 6722 instances per slide. Virtual positive slides are generated by randomly replacing a portion of the negative instances with positive instances. Seven different ranges of replacement quantities are defined, representing the upper and lower limits of the number of replaced instances for each experimental group. Within each range, the number of replaced instances in a virtual positive slide is randomly generated from a uniform distribution. The training set includes 400 slides, with 200 negative slides and 200 randomly selected slides used to generate virtual positive slides. The test set includes 200 negative slides, with 100 negative slides and 100 randomly selected slides used to generate virtual positive slides. We compared the classification performance of CLAM, transMIL, and our proposed method on the virtual dataset as shown in Response Fig. 13.

CLAM and transMIL utilize features extracted from all instances as slide-level features and classify slides using cross-entropy loss. The feature extraction model typically is a model pretrained on ImageNet. CLAM accelerates model convergence by introducing a clustering-based constraint loss at the instance level, which enforces significant differences between high-scoring and low-scoring instance features within a positive slide. CLAM assumes that the label of the instances with the high probability of being positive on the slide is consistent with the label of the entire slide. The larger the proportion of positive instances, the higher the probability that the two labels will be consistent. The convergence speed improves with a higher proportion of positive instances.

For transMIL, there is no instance selection or constraint mechanism. As a result, the model experiences more fluctuations during the training process. As the proportion of positive instances increases, the specificity between positive and negative feature sequences also increases. Consequently, the model converges faster and achieves better performance on datasets with a higher proportion of positive instances.

In contrast, our method introduces supervisory information during the feature extraction phase, enhancing the specificity of whole-slide representations. This allows the model to converge quickly even when the slides contain only sparse positive instances, resulting in the highest test set AUC.

It is important to note that the virtual dataset constructed here is significantly simpler than a real-world dataset, as the positive instances and negative slides used may exhibit notable differences in staining style or other latent features. This reduces the difficulty of identifying sparse positive instances. Although CLAM and transMIL perform well on virtual data, when applied to real data, the best result for the comparative methods on the internal dataset is 0.654, and that on the external test set is 0.504. In conclusion, these methods are not suitable for cytopathology data with sparse positive cells and limited image quality. In contrast, our method proves to be effective on real data.

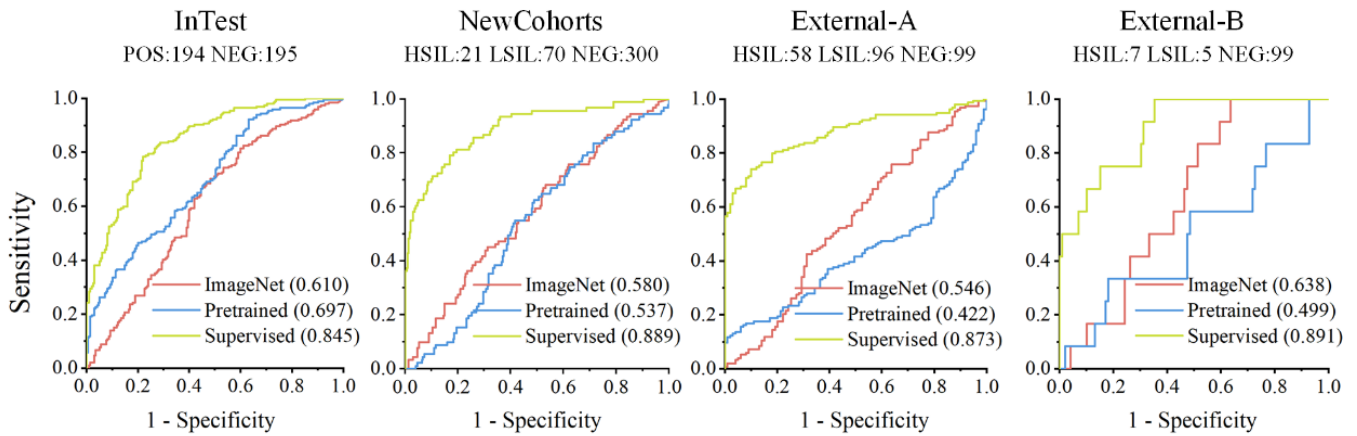
6. I want to know why the accuracy of Figure 5 and Figure 4 is different, I mean if I understand correctly, the results of the “supervised” in Figure 5 should be the same as “ours” in Figure 4, but they are different in the paper.

[Response 3-6] Thanks for the suggestion. Our previous description is not clear enough. This difference is due to the fact that the second stage of the model in Figure 4 (Figure 3 in revised version) experiment uses data augmentation, whereas Figure 5 experiment dose not.

During the training process of the model in the second stage, data augmentation can be applied. We first inferred all instances without augmentation using the model from the first stage and recorded the top k instances within the slide. Subsequently, we performed data augmentation on the selected instances before extracting the features, thus indirectly enhancing the input to the model in the second stage.

We repeated the experiment in Figure 5. All comparison methods employed data augmentation strategy. As shown in Response Fig. 14, compared to the original result, new result does not lead to any change in the conclusions (supervised representation is more effective and robust compared to self-supervised representation and ImageNet pretrained representation).

The results also indicate that data augmentation makes it more difficult to fit model on our dataset. When data augmentation is applied, there is a decrease in the AUC for identifying positive slides in the internal dataset. However, on the external dataset, data augmentation yields better results. On NewCohort and External-B, our method's AUC increased by 0.050 and 0.047 respectively, while the AUC decreased by 0.031 on External-A.



Response Fig. 14 | The ROC curves of Supervised, Pretrained and Pretrained on ImageNet (ImageNet for short in legend) features on InTest, NewCohorts, External-A and External-B sets. AUC and sample size are displayed in the legends and above figures respectively. All methods use data augmentation.

7. The manuscript needs to be polished severely. Plenty of grammar errors can be found.

[Response 3-7] Thanks to the suggestion. We have made our best effort to revise areas of poor language throughout the manuscript.

Response to Reviewer #4: (Remarks on code availability)

The codes are fine with references.

Response to Reviewer #5: (Expert in imaging and microscopy, engineering, cancer screening, and health disparities)

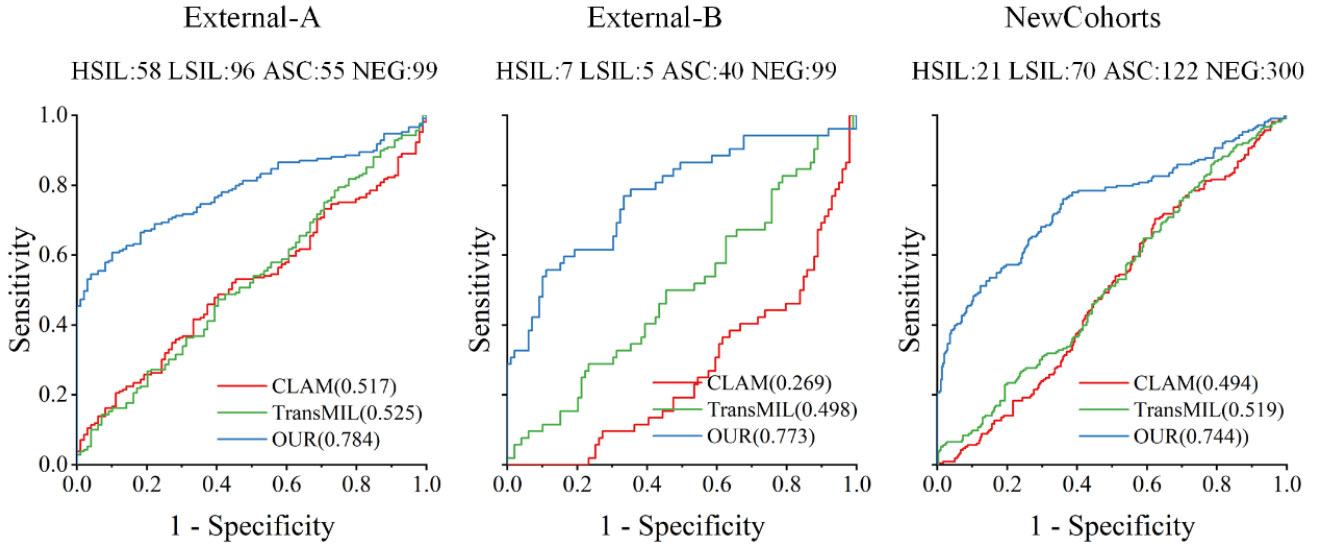
Bai et al. demonstrated a low-cost compact microscope and an AI classification algorithm for cervical cytology to enable cancer screening in resource-limited settings. The prototype microscope captured similar cell morphology features compared to the standard microscope but with reduced cost and size. The authors developed a two-stage classification algorithm based on a Transformer model, which was then trained and validated using microscopic images of cytology slides captured by the proposed compact microscope. However, I cannot recommend this manuscript to be published in Nature Communications due to several significant drawbacks in this paper. To validate the clinical performance of the proposed method, a more comprehensive analysis of the testing accuracy is necessary. In addition, significant improvements to the system setup and imaging procedure are also required to facilitate its application in various clinical settings, especially in resource-limited areas. My comments are listed below.

Major comments:

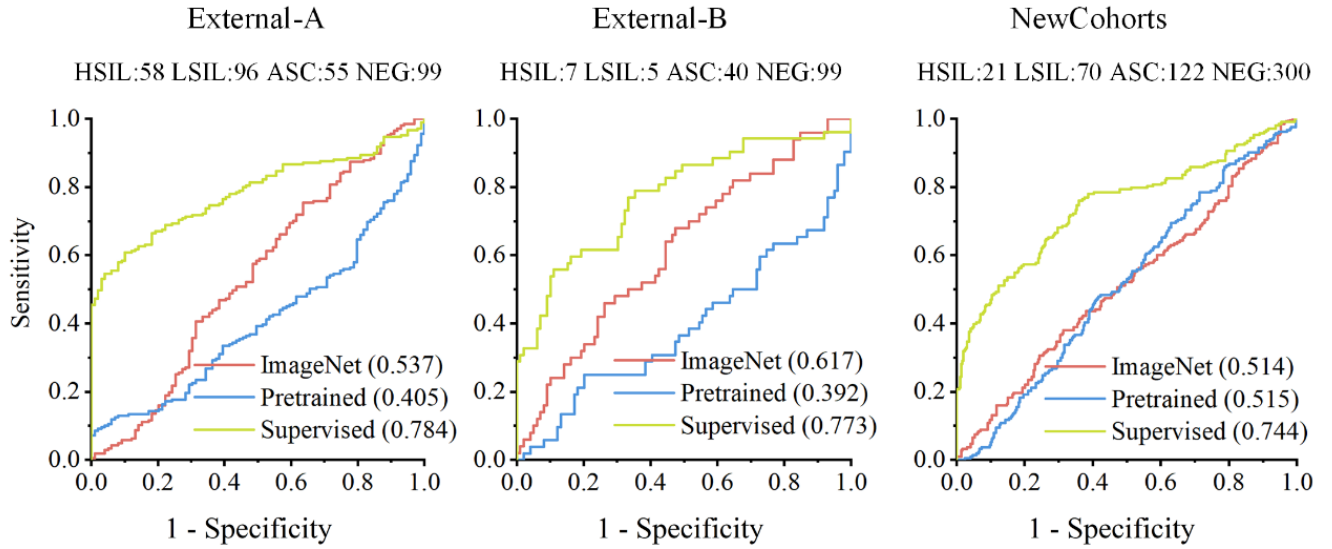
1. The authors evaluated the performance of the proposed method and the screening accuracy is shown in Figure 4. The authors make the argument that the model can generalize well across datasets. However, when testing the system using the two external datasets and NewCohorts dataset, the authors exclude atypical squamous cells (ASC) cases because of the lower classification performance. ASC is the most common abnormal finding in cervical cytology, which can also be reflected in the number of ASC cases within the patient cohort. In addition, ASC cases were included in the training dataset. Given these reasons, the authors should include ASC cases in all cases when evaluating the system performance, including Figs. 4, 5 and 7. The sensitivity and specificity considering all positive cases should be evaluated and reported. The authors should further optimize the hardware or the software of the proposed system if the accuracy is compromised. **[Response 5-1]** We thank the referee for this critical feedback. We re-plotted the figures to show the results considering ASC, where Response Fig. 15, 16, and 17 correspond to Figure 4 (Figure 3 in revised version), 5, and 7, respectively. We observed a decline in the results of all experiments after adding ASC. First, the AUC of our method decreased by 0.145, 0.088, and 0.118 on three external test sets as shown in Response Fig. 15. Secondly, in the ablation experiments on the feature extraction process in the first stage and the classification model in the second stage, we observed similar conclusions as shown in Response Fig. 16 and Response Fig. 17. Nevertheless, on the same test sets, our method still outperforms the other methods.

As a diagnostic category characterized by atypical or minimal lesions, ASC (Atypical Squamous Cells) presents inherent challenges for accurate identification compared to more well-defined pathological categories such as HSIL (High-grade Squamous Intraepithelial Lesions) and LSIL (Low-grade Squamous Intraepithelial Lesions). In clinical practice, significant interobserver variability has been observed among pathologists when diagnosing ASC cases. From the perspective of precancerous lesion screening, ASC may indicate earlier or atypical pathological changes. In regions where cervical cancer screening programs are well-established and the cervical cancer incidence is relatively low, ASC could represent an important focus for current research. However, in resource-limited areas with insufficient screening programs and higher cervical cancer incidence rates, prioritizing the identification of well-defined cytological abnormalities such as HSIL and LSIL holds greater significance. In such cases, referring high-risk individuals to centralized healthcare facilities for further diagnosis and treatment is a more practical and urgent strategy. Overall, our approach prioritizes screening accessibility over maximum performance while ensuring the effectiveness of screening for high-risk populations, striking a balance between screening availability and diagnostic accuracy.

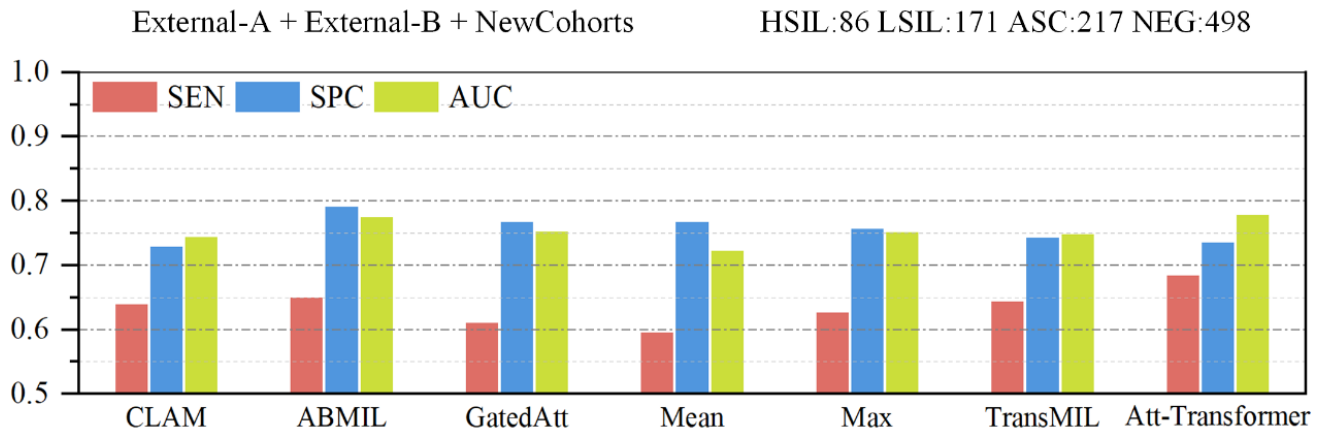
Our paper focuses on the issue of screening high-risk populations for cervical cancer in resource-limited areas. Considering that accurate identification of ASC types is not the central focus of our paper, we decide to include the results related to the ASC category in the appendix.



Response Fig. 15 | The figure illustrates the ROC curves of CLAM, TransMIL and our methods on NewCohorts, External-A and External-B sets. AUC and sample size are displayed in the legends and above figures respectively.



Response Fig. 16 | The ROC curves of Supervised, Pretrained, and Pretrained on ImageNet (ImageNet for short in legend) features on NewCohorts, External-A, and External-B set. AUC and sample size are added to the legend and above the figure, respectively.



Response Fig. 17 | AUC, sensitivity and specificity of different methods: CALM, ABMIL, GatedAtt, MeanPool, MaxPool, TransMIL and Att-Transformer. Dataset size is displayed above the figure.

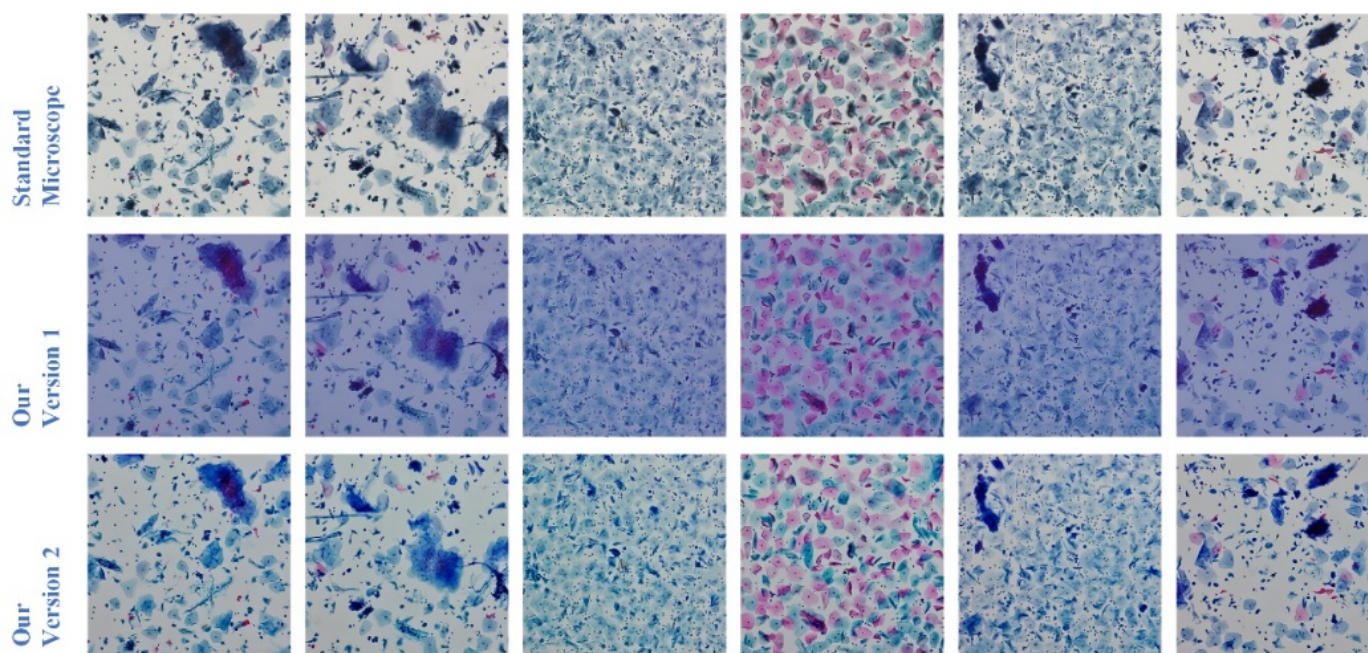
2. As shown in Fig. 4, the sensitivity and specificity of the proposed method varied a lot when using different datasets. Could the authors comment on that?

[Response 5-2] Unlike pathologists interpreting slides, where each slide's result is definitive, a model outputs a predicted probability of positivity for each slide. Typically, we choose a probability threshold to determine the boundary for classifying a slide as positive. Sensitivity and specificity change with variations in the threshold. The three external test sets come from different sources, so there are differences in the data distribution. For example, when the positive slides are relatively typical, the predicted values of positive slides by the model may be concentrated at higher scores. However, when the positive slides are relatively atypical, the predicted values of positive slides by the model may have a wider distribution range. In our experiments, using a threshold of 0.5 as the decision boundary for classifying the model's predictions as positive or negative does not guarantee that sensitivity and specificity are balanced across all test sets. Sensitivity and specificity are a pair of antagonistic evaluation parameters. When considering the AUC metric, which summarizes the model's performance across different thresholds, the model demonstrates relatively consistent performance across the three external test sets.

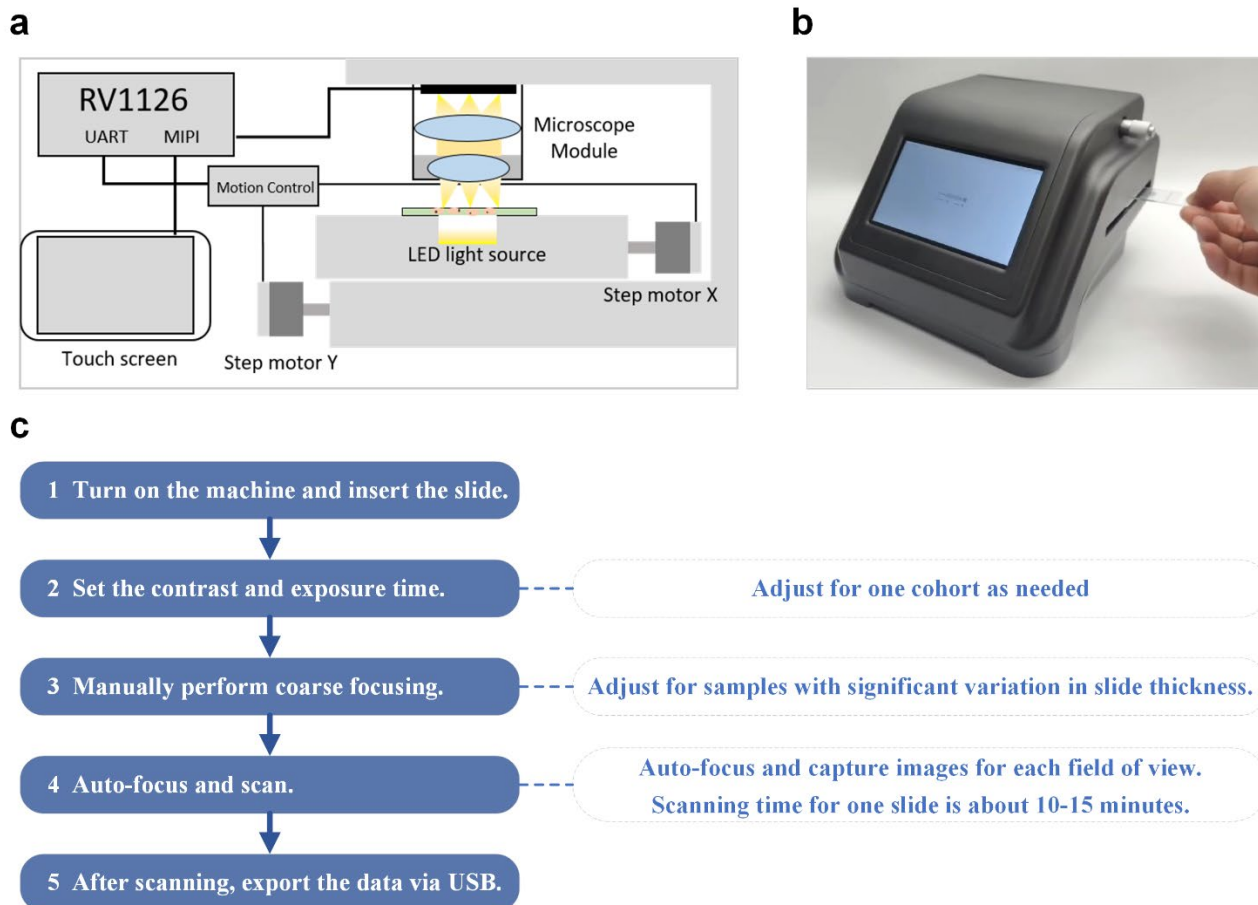
3. How are the classification metrics calculated? Is this done at WSI level or individual tile level? How do you account for the redundant images? The authors should detail this in the methods section.

[Response 5-3] In our dataset, each whole slide image represents a single patient, with one slide per patient. This setup differs from the scenario in histopathology, where a single patient is often associated with multiple slides.

4. Although the authors proposed a compact microscope that can be used in resource limited settings, only a rough, incomplete system prototype was demonstrated in this study, which still requires significant optimization to improve usability of the device in the clinic. The image acquisition relied on manual movement of the slides and capturing around 90 images to cover the entire field of view, which is highly labor-intensive and unreliable. The authors should automate the image acquisition and analysis procedure to enable the clinical application of the system, or the authors should acknowledge this limitation in the manuscript. The system packaging should also be further optimized. It would be helpful to include a clear description or flow chart detailing which portions of the image acquisition are manual and which are automated.



Response Fig. 18 | Imaging results of the same target using a standard microscope, our first version microscope, and our second version microscope.



Response Fig. 19 | a) Schematic diagram of our compact imaging system. b) Appearance of our compact imaging system. c) Whole slide image acquisition flowchart with descriptions of each step.

[Response 5-4] Thanks for the suggestions. Due to the long duration of our overall research cycle, we prioritized validating our algorithm on a prototype during the hardware development phase. We implemented several optimizations as shown in Response Fig. 19. We first added an automatic scanning system, enabling the device to autonomously complete the slide scanning process. Additionally, we introduced a control and monitoring panel, which allows for convenient adjustment of various imaging parameters, such as exposure time and contrast, while also facilitating real-time monitoring of the imaging results. Lastly, we improved the system packaging by integrating a battery and power interface, as well as adding USB control and output interfaces. We also enhanced the imaging quality of the system. Various imaging metrics and styles were compared, as illustrated in the Response table 2 and Response Fig. 18. Despite these improvements, the overall hardware cost remains below \$500. We have supplemented a video to visually demonstrate our system and the scanning process. Details are outlined below:

For the automation of slide data acquisition, we selected two sets of 20mm stepping screws arranged in an orthogonal combination to construct a two-dimensional translation stage, which works in conjunction with the Voice Coil Motor (VCM) of the microscopic imaging system for autofocus. With an LED scatter-type backlight board as the illumination source, we can achieve fully automated cell data acquisition from 10 to 15 minutes depending on the areas of slide.

To facilitate portable section scanning and image observation, we chose a main control chip, RV1126, which supports the Linux operation system, and equipped it with a 10-inch screen. On the entire embedded platform, we developed an interactive scanning software based on Qt. The power consumption of the entire system is controlled within 4W, and the device is configured with a 6000mAh battery, enabling continuous operation for over 6 hours.

To obtain more structural characteristics of cervical cells, we have updated the microscopic imaging system with enhanced imaging performance. The system utilizes two sets of aspherical lens designed objectives and tube lenses to

construct the entire optical system, with the focal lengths of the tube lens and objective lens being 3.6mm and 4.7mm respectively, forming an approximately 3.1x magnification system. The imaging system utilizes the Sony IMX415 image sensor, which features a 1/2.8-inch optical format and a pixel pitch of 1.45 μm . This setup allows the system to achieve a pixel size of 0.46 μm , enabling high-resolution imaging, which is suitable for detailed analysis. The module also integrates a VCM autofocus module with a travel range of 200 micrometers and a step precision of 3 to 8 $\mu\text{m}/\text{mA}$, eliminating the need for complex focusing structure.

Overall, the improved system has achieved a comprehensive enhancement in usability, making it well-suited for screening in resource-limited settings. We have used the upgraded system to scan 570 slides for a sub-study on HPV-positive data, addressing other reviewers' concerns about our work (see response 6-3). On the HPV-positive dataset, our method achieved an AUC of 0.85, demonstrating that the algorithm developed based on the prototype remains well-aligned with the improved system.

5. Compared to a standard microscope, the images acquired using the proposed compact microscope exhibit a different color tone with a pronounced purple/blue background across the entire field of view. Could the authors provide an explanation for this? Has any white balance correction been explored? While we understand that the algorithm may be able to compensate for this if it is only trained on images from this microscope, any real world implementation of the system should allow for the visual review of images by a pathologist when a pathologist is available, and the while we understand there may be some artifacts due to the low-cost nature of the microscope, it seems that a white balance correction should be implemented.

[Response 5-5] Thanks for the suggestions. The color abnormalities in the images are caused by incorrect parameter settings in the automatic white balance correction method. To address this issue, we fine-tuned the parameters and resolved these defects using an improved device.

Response table 2. metrics for image quality			
metric	Standard microscope (HR)	Our version 1 (v1)	Our version 2 (v2)
PSNR (with HR)	\	15.76	19.89
SSIM (with HR)	\	0.497	0.644
Noise Std Dev	5.49	19.63	4.24
Edge Variance	218.17	33.72	16.82
percentage of lost focus ¹²	0.013	0.017	0.015
Saturation Probability	0.000374	0.000000	0.000115
Resolution (1951 USAF)	912.3 lp/mm (0.54 μm)	287.4 lp/mm (1.74 μm)	574.7lp/mm (0.87 μm)

We consulted pathology experts regarding the suitability of the imaging style of the new system and received positive feedback. Response Fig. 18 shows the imaging results of the same target captured by a standard microscope, the original microscope and the improved microscope. It can be observed that the imaging style of the new system is relatively consistent with that of the standard microscope. Statistical metrics of the three imaging modes can be found in Response table 2. Additionally, the improved system includes a touch-screen interface that allows users to adjust exposure time and contrast interactively.

In fact, at the beginning of model training, we had already recognized the need for the model to adapt to different shooting conditions. First, we trained the model using data from multiple centers. Next, we employed a color-focused data augmentation strategy. This augmentation included color jittering in both the RGB and HSV color spaces, as well as enhancements in grayscale and contrast. Additionally, Gaussian blur augmentation was applied. Data collected with the optimized equipment was then used to further test the model and we found the algorithm developed based on the prototype remains well-aligned with the improved system.

6. The authors demonstrate a lower contrast and higher noise level of the proposed compact microscope compared to the standard microscope. However, instead of qualitative comparisons of the image quality of the 100 image pairs, the authors should provide a comprehensive quantitative analysis to compare the two imaging systems.

[Response 5-6] Thanks for reviewer's valuable suggestion. As a similar question has already been raised by a previous reviewer, I will use the same response to address this concern, hoping to clarify the reviewer's doubt. **See response 1-7.**

Minor points: reviewer #5

1. The comparisons between cytology images and histopathology images shown in Fig. 1c and 1d are distinct from the other panels in Fig. 1. The authors might consider moving Fig. 1c and 1d to a different figure or supplementary materials. Consider removing Figure 1d entirely, as it does not contribute to the focus of your paper.

Re: Thanks. We have taken your suggestion into account and moved the relevant content to the appendix.

2. For the input of the first stage of the network, why is the image patch resized to a smaller size?

Re: Thanks. 224 pixel size is the most common input format for transformer models, as it conveniently divides an image into 14×14 regions, each with a size of 16 pixels, where each region corresponds to a token in the input sequence. In this study, we adopt the simplest approach to align the size of our image with this format.

3. When dividing a large image to smaller 256×256 image patches, what is the size of the overlapped area between two neighboring patches? Could the authors explain why a redundant sampling is needed, and how is the size of the overlapped area determined?

Re: Thanks for the suggestions. The redundancy between each pair of neighboring patches is 64 pixels in width, which corresponds to a physical size of 55.68 micrometers. Using a redundant sliding window ensures that a cell (typically smaller than 55.68 micrometers) appears completely at least once within the image, rather than being split at the boundary. A complete cell structure provides the model with more comprehensive information, which enhances its ability to make accurate distinctions.

4. How many patients are included for each dataset used in this study?

Re: In contrast to the scenario in histopathology where a patient may correspond to multiple slides, in our dataset, each slide corresponds to a single patient. A total of 3,899 participants were involved in training, selecting, and conducting internal testing of the model. Additionally, 972 generic participants and 570 participants who have been tested as HPV positive were included in the external testing of the model. Overall, a total of 5441 participants took part in the experiment.

5. It is unclear whether inference was run in a real-time or near real-time manner with images collected with the microscope system. What are the barriers here?

Re: In fact, the inference time primarily depends on the sample area within the slide and the performance of the hardware used. Additionally, the large volume of collected sample data can be sent to high-performance computing devices for rapid, parallel inference. The minimum computational resources required for inference are very modest, enabling completion even on devices equipped with only a CPU and no GPU. When test on a server which equipped with one NVIDIA TITAN V graphics processing unit (GPU) and Intel(R) Core (TM) i9-9940X CPU @ 3.30GHz, the inference time for a single slide is within 3 minutes. Currently, the computation time for inference is significantly shorter than the time required for slide scanning. This means that while the next slide is being scanned, there is ample time to complete the inference for the previous slide.

6. The paper should be proofread to correct many grammatical errors, typos, and general edits. A few are noted below, however this is not a comprehensive list:

- Define what negative and positive cases are in the introduction.
- Typo in "sparsely-distributed lesion cells" found throughout the manuscript.
- In Figure 7, the legend should be "HSIL" instead of "HISL". Also CLAM is mentioned twice in the caption.

d. In the abstract, "for high-risk population" should read "for high-risk populations"

e. On page 3, "easy of connection" should read "ease of connection"

Re: We sincerely appreciate the reviewer's comments, and we have made the corresponding revisions in the relevant sections.

7. It would be helpful to consistently refer to the new method as either "OUR" or "Att-Transformer" in the figures. The use of "OUR" is somewhat confusing as it can appear to be an undefined acronym.

Re: Thanks. We now use DualCytoNet to replace OUR.

8. The Figure 3 caption needs to be expanded to include additional details. For example, an explanation of the term "hard samples" would be helpful.

Re: Thanks. We now add more information in Figure 3 (Figure 4 in revised version) and further explain "hard samples".

9. The term "grassroots" is a bit unorthodox to describe a low resource setting. The authors may want to reconsider this term.

Re: Thanks. We now use "resource-limited area" instead of "grassroots area".

Response to Reviewer #6: (Expert in cervical cancer screening)

This is a manuscript investigating AI supported cervical cytology for intended use in low-resource settings. Whereas the goal of the research is highly commendable, and the work is very timely and interesting, I have the following major comments for consideration.

1 The Introduction states in lines 49-57, that “cytology screening remains crucial for reducing the incidence of cervical cancer”. This completely ignores the possibility of HPV-based screening, with or without the addition of self-sampling in a low-cost medium, which is a high throughput technology at low cost. Substantial efforts worldwide have been dedicated to push the cost of cervical screening down through clever use of HPV testing to rule out women with no risk for cancer. And it is surprising to see no acknowledgment or position from the authors on this possibility, especially as they allude to the WHO elimination goal on row 134.

[Response 6-1] Thanks for the reviewer's comments. We acknowledge the missing discussion regarding the HPV-based screening strategy in our manuscript, and we have added this section in Introduction and Discussion. The HPV-based screening, as a method prioritized by WHO, appears to be a suitable screening approach for resource-limited regions, given its high-throughput capability, compatibility with self-sampling, and independence from pathology experts.

In the introduction: *“In recent years, world health organization (WHO) has recommended HPV screening for high-throughput screening of cervical cancer^{16,17}, and extensive work has been carried out¹⁸⁻²¹. However, HPV test usually has the problem of relatively low specificity²²⁻²⁴. Cytopathology-based screening can directly provide information on cellular morphological changes and is usually a necessary step following a positive HPV result for the final diagnosis of cervical precancer detection.”*

In discussion: *“As WHO has recommended HPV screening for high-throughput screening of cervical cancer, joint screening with HPV test is an important potential application of our system. We tested our method on a dataset (n=570) that were already identified as HPV-positive (supplementary Fig. 12). Our overall AUC was 0.850 and 0.755 when excluding ASC or not. Our method is still able to effectively distinguish SIL in HPV-positive slides. Considering the relatively low specificity of HPV test, our model can effectively exclude a large number of cytopathology negative cases, thereby improving the specificity of co-testing. These results indicate that our system still can play an important role in the co-test or HPV-first screening scenarios.”*

2 The problem of cervical cytology is that of subjectivity in diagnosis and particularly, false negatives, and correct training of the algorithm is essential, as well as maintenance. The European Union will shortly advise against the use of cytology instead of focusing on the importance of introducing HPV-screening for the good of the female population. I am surprised at the lack of consideration in the paper on how the authors defined and validated the gold standard diagnoses of subclasses ASC, LSIL and HSIL, especially since ASC suffers from notorious inter-observer variation. (AI could hopefully improve substantially upon this situation, indeed, but right now I need to understand better how the algorithm was basically validated as correct.)

[Response 6-2] We fully agree with this referee on the importance of introducing HPV-screening for the good of the female populations. One advantage of HPV screening is its ability to avoid observer bias, which is a significant challenge in cytopathology for precancerous screening. Machine learning algorithms are objective and stable, mainly relying on the datasets used for their training. From a machine learning perspective, if there is a sufficiently large and accurate enough dataset, it has the opportunity to train a model that not only possesses strong diagnostic capabilities but also ensures the stability of its standards. This is a task that requires the collective effort of the entire field.

In order to avoid deviation in the process of manual labeling, two groups of pathology experts were involved in annotating our dataset, each consisting of two junior pathologists and one senior pathologist. The two junior pathologists

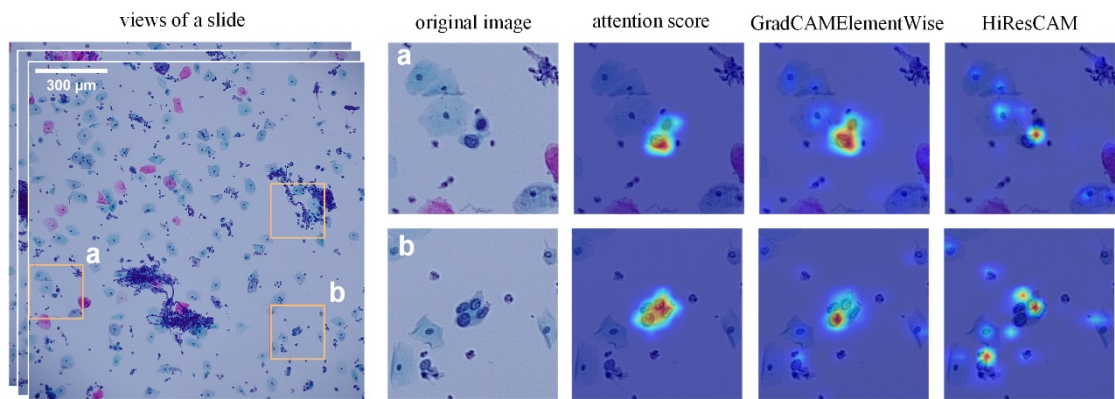
independently completed the slide annotations, and in cases where their annotations differed, the senior pathologist reviewed and resolved the discrepancies. We abandoned questionable annotations. To ensure the reliability of annotations, all slides were scanned twice: once with a standard microscope and once with our portable microscope. Pathology experts annotated the slides obtained from the standard microscope, and these annotations were subsequently transferred to the corresponding slides acquired by the portable microscope.

We made our best efforts to collect a diverse and substantial dataset to support our study. A total of 3,135 slides from four sources were used to train the model, and 375 slides and 389 slides were served as the validation and internal test sets respectively. To better evaluate our model, we tested it on four external test sets: 513 slides from the same source as the training set but from a different time period; 308 and 151 slides collected from two other primary hospitals; 570 slides from HPV-positive cases. Overall, a total of 5441 slides are used in our experiment. More details can be found in supplementary table 3, supplementary table 4 and supplementary table 5.

Our model achieved AUCs of 0.889, 0.873, 0.891 for detecting SIL from negative slides by our model on the three external test sets and achieved AUC of 0.85 on HPV-positive test set. These results demonstrate the model's stable performance on external datasets and its effectiveness in identifying SIL types.

To further validate our model, we make use of the attention scores output by our model to generate Attention Score Map. We also generated CAMs based on the feature maps from the penultimate multi-head self-attention (MHSA) layer. We chose GradCAMElementWise and HiResCAM⁸ to plot the map as shown in Response Fig. 20. The ROIs identified by the heatmaps closely align with the ROIs identified from a pathological perspective, indicating that the lesion areas emphasized by our method contain reliable semantic information.

Overall, based on our careful annotation process, a considerable amount of training data, and the effective results obtained by the model on multiple external test sets along with various visualization outcomes, our algorithm was essentially validated as correct.



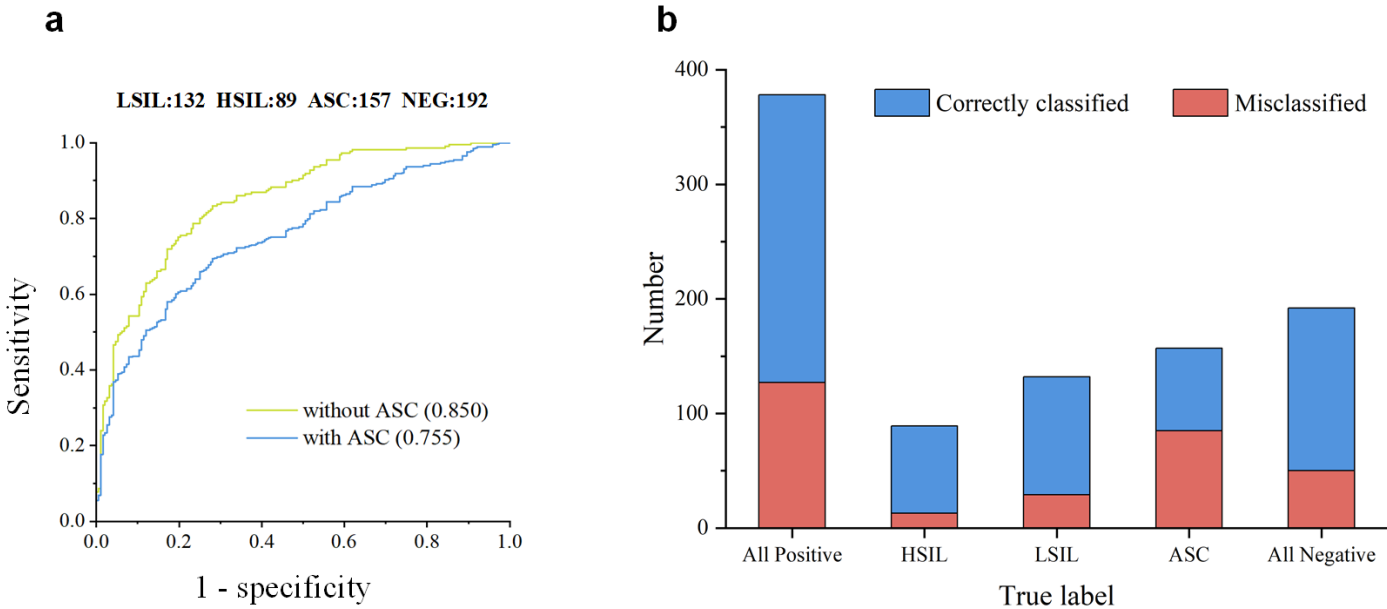
Response Fig. 20 | Visualization of attention mechanism and interpretability. Cell-level ROI and instance-level ROI are acquired from the attention score of the first and second stage models. We frame out instances with the top 200 positive probabilities as instance-level ROI from a view of a slide and plot the heat map of our attention score in stage one. Additionally, class activation maps (CAMs) generated from the output of the penultimate multi-head self-attention layer of our model are plotted, specifically for positive class. Two types of CAM are shown: GradCAMElementWise⁷ and HiResCAM⁸.

3 That said, the large remaining problem is how to handle triage of HPV-positive women to the limited resources of colposcopy and biopsy. Here, cytology of HPV-positive smears with AI-assisted efforts could be useful and I was somewhat surprised to see that this was not the focus of the study. For Nature group to consider this manuscript further, it is my opinion that the authors must conduct a substudy in HPV-positive samples to present a more timely case compatible with present global algorithms for more regions than those in China.

[Response 6-3] Thanks for the suggestions. In the context of HPV as the primary screening strategy, we tested our method on a dataset that was already identified as HPV-positive. The results are shown in the Response Fig. 21. Our overall AUC was 0.76 when including the ASC category and 0.85 when excluding ASC. Our method is still able to

effectively distinguish the cytological categories in HPV-positive slides. We also analyzed the misclassification patterns across different categories. The results indicate that the model's diagnostic performance was better for the HSIL category compared to LSIL and ASC, with the lowest performance observed in the ASC category.

In combined screening, our model can effectively exclude a large number of cytopathology negative cases, thereby improving the specificity of co-testing. For instance, we can recommend patients who are HPV-positive and have the highest scores of our algorithm for biopsy, while those with relatively lower algorithm scores but still HPV-positive can be referred for endoscopic examination.



Response Fig. 21 | a) The ROC curve of the model on HPV-positive cases, with an AUC value of 0.76 when including the ASC category and an AUC value of 0.85 when excluding it. The test set contains 89 HSIL, 132 LSIL, 157 ASC, and 192 NEG cases. b) The model's classification results for each subclass.

4 I entirely miss reports on which set of pathologists/panels ultimately classified and validated what constituted an abnormal cell, or a true lesion apparently defined as “subclasses ASC, HSIL, LSIL, NEG” - ASC and LSIL in particular are notoriously subjective reference standard issues. Row 454 should state ASC, LSIL, HSIL, NEG, by the way.

[Response 6-4] Thanks, we recognize that we did not give a clear definition of the cell classes in our original manuscript. The annotations are based on the Bethesda System (TBS). In our study, we focus on the screening of intraepithelial lesion which mainly consists of ASC, LSIL and HSIL samples, while negative samples are the same as negative for intraepithelial lesion or malignancy (NILM) samples. However, the interpretation process inevitably introduces subjective factors, which are difficult to quantify. We adopted the following annotation process to ensure the validity of the annotations as much as possible:

Two groups of pathology experts were involved in annotating our dataset, each consisting of two junior pathologists and one senior pathologist. The two junior pathologists independently completed the slide annotations, and in cases where their annotations differed, the senior pathologist reviewed and resolved the discrepancies. We abandoned questionable annotations. To ensure the reliability of annotations, all slides were scanned twice: once with a standard microscope and once with our portable microscope. Pathology experts annotated the slides obtained from the standard microscope, and these annotations were subsequently transferred to the corresponding slides acquired by the portable microscope.

5 It is not surprising that the AUC is 0.89 on NewCohorts as the algorithm essentially returns to the original 4 centres' type of classification of subclasses - is there some form of circular over-fitting going on here? Overall I find an AUC of 0.90 to be very optimistic, especially given that cytology diagnosis of underlying histological lesions is not a very precise science to begin with.

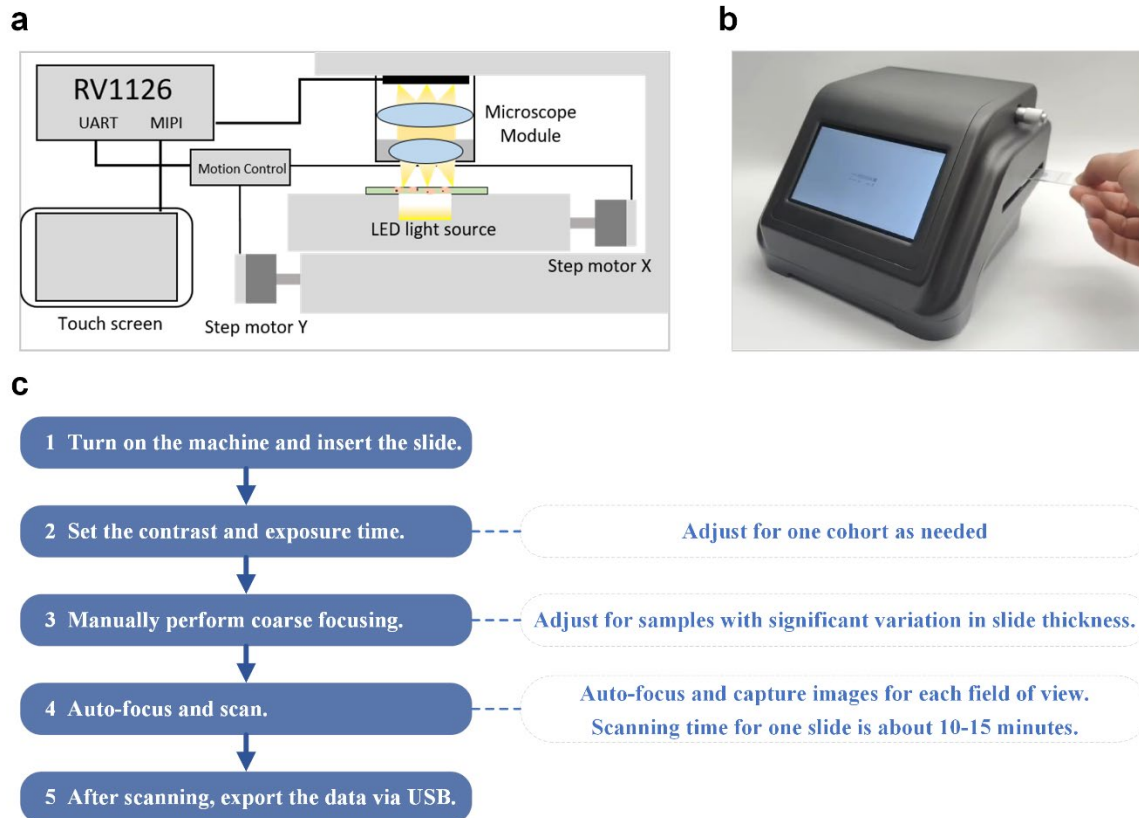
[Response 6-5] We thank the referee for raising this point. On three external test sets, our method achieved AUCs of 0.889, 0.873, and 0.891, with an overall AUC of 0.9 for detecting SIL from negative slides. When including the ASC category, our method achieved AUCs of 0.784, 0.773, and 0.744 respectively. Our method achieves a relatively high AUC for detecting typical lesions such as SIL, but when considering the more challenging ASC category, there is still significant room for improvement.

From a machine learning perspective, overfitting refers to the phenomenon of excessively fitting the training data. Overfitting typically shows significant discrepancy in model performance between the training and test sets, or a large difference between the internal and external test sets.

In the experiments identifying SIL on three external test sets, our method achieved AUCs of 0.889, 0.873, and 0.891, with an overall AUC of 0.9. The close AUC values across the three external test sets indicate that overfitting was minimal. NewCohorts is a test set similar to the internal test set, as it shares the same data source but was collected during a different period and differs from the other two datasets. This indicates that our classifier is more robust. Besides, on the HPV-positive test set, our method still achieved an AUC of 0.85 in identifying SIL. This also demonstrates the robustness of our method and suggests that our approach does not exhibit significant overfitting.

6 Ideally, a fully external evaluation of the tiny microscope and classifier should be done by a collaborating partner in a clinic/pathology laboratory reference cytology of very high quality, for example in the US or Europe.

[Response 6-6] Our test data included slides from laboratory reference cytology of very high quality in China. Our system accurately screens for SIL types, demonstrating its effectiveness. We have developed our hardware into a highly usable product, as shown in Response Figure 22. Additionally, we plan to open-source our model architecture and weights, enabling further validation of our system in other laboratories. We also plan to conduct additional validation studies in the US or Europe.



Response Fig. 22 | a) Schematic diagram of our compact imaging system. b) Appearance of our compact imaging system. c) Whole slide image acquisition flowchart with descriptions of each step

Minor points: reviewer #6

1. I went to the authors' github account to try to look at the Demonstration data alluded to in row 580 but could not immediately locate that file. Please clarify exactly where in the github material this is located.

Re: The original link may not be very noticeable. We will update the webpage content as soon as possible. Please refer to the following link. Open source model weights at <https://huggingface.co/BruceAwake/high-risk-Cervical-Precancerous-Screening/tree/main/weights> and LRWSI demo file at https://huggingface.co/BruceAwake/high-risk-Cervical-Precancerous-Screening/tree/main/WSI_DEMO

2. "Grassroots centers" - what is meant by this term exactly?

Re: We now use "resource-limited area" instead of "grassroots area".

3. Row 286-287 typos, it should be HSIL and LSIL.

Re: Thanks. We have corrected the error.

Reference

1. Xu, H. *et al.* A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
2. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
3. Han, X. *et al.* Pre-trained models: Past, present and future. *AI Open* **2**, 225–250 (2021).
4. He, K. *et al.* Masked Autoencoders Are Scalable Vision Learners. Preprint at <https://doi.org/10.48550/arXiv.2111.06377> (2021).
5. Dosovitskiy, A. *et al.* An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Preprint at <http://arxiv.org/abs/2010.11929> (2021).
6. Dauphin, Y. N., Fan, A., Auli, M. & Grangier, D. Language Modeling with Gated Convolutional Networks. Preprint at <https://doi.org/10.48550/arXiv.1612.08083> (2017).
7. Gildenblat, J. & contributors. PyTorch library for CAM methods. GitHub. <https://github.com/jacobgil/pytorch-grad-cam> (accessed Jan 25, 2025).
8. Draelos, R. L. & Carin, L. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. Preprint at <https://doi.org/10.48550/arXiv.2011.08891> (2021).
9. Cheng, S. *et al.* Robust whole slide image analysis for cervical cancer screening using deep learning. *Nat. Commun.* **12**, 5639 (2021).
10. Cao, M. *et al.* Patch-to-Sample Reasoning for Cervical Cancer Screening of Whole Slide Image. *IEEE Trans. Artif. Intell.* 1–11 (2024) doi:10.1109/TAI.2023.3323637.
11. Lin, H. *et al.* Dual-path network with synergistic grouping loss and evidence driven risk stratification for whole slide cervical image analysis. *Med. Image Anal.* **69**, 101955 (2021).
12. Chen, D.-J., Chen, H.-T. & Chang, L.-W. Fast defocus map estimation. in *2016 IEEE International Conference on Image Processing (ICIP)* 3962–3966 (IEEE, Phoenix, AZ, USA, 2016). doi:10.1109/ICIP.2016.7533103.
13. Cheng, S. *et al.* Robust whole slide image analysis for cervical cancer screening using deep learning. *Nat. Commun.* **12**,

- 5639 (2021).
14. Shao, Z. *et al.* TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. Preprint at <https://doi.org/10.48550/arXiv.2106.00908> (2021).
15. Lu, M. Y. *et al.* Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat. Biomed. Eng.* **5**, 555–570 (2021).
16. Canfell, K. Towards the global elimination of cervical cancer. *Papillomavirus Res.* **8**, 100170 (2019).
17. Shastri, S. S. *et al.* Secondary Prevention of Cervical Cancer: ASCO Resource–Stratified Guideline Update. *JCO Glob. Oncol.* e2200217 (2022) doi:10.1200/GO.22.00217.
18. de Sanjose, S. & Holme, F. What is needed now for successful scale-up of screening? *Papillomavirus Res.* **7**, 173–175 (2019).
19. Hu, L. *et al.* An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. *JNCI J. Natl. Cancer Inst.* **111**, 923–932 (2019).
20. Serrano, B. *et al.* Worldwide use of HPV self-sampling for cervical cancer screening. *Prev. Med.* **154**, 106900 (2022).
21. Cox, B. & Sneyd, M. J. HPV screening, invasive cervical cancer and screening policy in Australia. *J. Am. Soc. Cytopathol.* **7**, 292–299 (2018).
22. Arbyn, M. *et al.* Chapter 9: Clinical applications of HPV testing: A summary of meta-analyses. *Vaccine* **24**, S78–S89 (2006).
23. Denton, K. J. *et al.* The Sensitivity and Specificity of p16INK4a Cytology vs HPV Testing for Detecting High-Grade Cervical Disease in the Triage of ASC-US and LSIL Pap Cytology Results. *Am. J. Clin. Pathol.* **134**, 12–21 (2010).
24. Benevolo, M. *et al.* Sensitivity, Specificity, and Clinical Value of Human Papillomavirus (HPV) E6/E7 mRNA Assay as a Triage Test for Cervical Cytology and HPV DNA Test. *J. Clin. Microbiol.* **49**, 2643–2650 (2020).

Response to reviewers

Dear Editors and Reviewers,

We sincerely appreciate the time and expertise invested by the reviewers in evaluating our manuscript. Their constructive comments have significantly improved the quality of our work. We are particularly grateful that the reviewers acknowledged the validity of our initial responses, which confirms our shared commitment to rigorous scientific standards.

Below, we provide a point-by-point response to each comment, incorporating all suggested revisions with corresponding modifications highlighted in the resubmitted manuscript.

Reviewer #1 (Remarks to the Author):

The revised manuscript has been substantially improved and presents a promising approach to cervical cytology precancerous screening in resource-limited regions by integrating a compact, cost-effective microscope with an AI-based classification system. The authors have provided detailed explanations in response to the reviewers' comments. The proposed two-stage classification model, which combines self-supervised learning with an attention-based transformer, is well-motivated and supported by small-scale multi-center validation. While the results demonstrate strong potential, several areas require minor further clarification and improvement.

One previous key concern raised previously was the resolution of the compact microscope, stated to be approximately 0.87 $\mu\text{m}/\text{pixel}$. The authors have addressed this issue by including visualized images and quantitative data.

I understand the rationale behind the two-step training approach, particularly with the support of the reply to the review comments. However, since readers may not have access to these responses, I suggest that the authors further elaborate on this idea within the manuscript.

The robustness of the model is discussed based on the given dataset, but additional field testing and validation may be necessary in the future, which may not be within the scope of this work.

The manuscript would benefit from professional editing to enhance readability and organization. Ensuring consistency between figures, text, and citations (e.g., missing references in citation lists) is essential for completeness.

Overall, this study presents an innovative AI-assisted cervical screening system with conceptual advancements aimed at addressing healthcare challenges in regions with limited medical resources. Additionally, some of the content from the "reply to reviewer" document provides valuable insights that could help readers better understand and appreciate this work. The authors may consider making their responses more concise and incorporating key points into the main text.

Recommendation: Accepted. Further review is not necessary.

[Response to reviewer #1] We sincerely appreciate your positive feedback on our initial responses. We have carefully incorporated all essential information into the main text and supplementary materials while maintaining the clarity and focus of the study's narrative. At this stage, we believe that introducing additional content to the main text may disrupt its coherence.

Regarding system robustness testing, we are actively pursuing opportunities to broaden the evaluation scope and include more diverse scenarios, and will publish the results in our future work.

Further, we have meticulously reviewed the manuscript line-by-line to resolve any stylistic or formatting inconsistencies, and have refined the text to ensure strict adherence to academic writing standards.

Minor points: reviewer #1

1 Figure quality needs improvement. Some figures could be reorganized and combined for clarity. The figure style should also be adjusted to align with Nature journal standards. Authors may also consider to improve the quality of the figure captions.

Re: We have revised the figure legends for multiple figures to ensure compliance with standard conventions and further optimized the image resolution for enhanced clarity.

2 Ensure consistent abbreviation of author names in the author contributions section. For example, Liu Xiuli -> X.L.L., but Shenghua Cheng -> S.C. Please review this section for consistency and clarity.

Re: The identified errors have been revised as suggested.

3 References [39] and [42] are duplicated. Please check the full reference list to eliminate such errors.

Re: We have thoroughly reviewed the reference list and corrected the identified errors.

3 All acronyms should be defined when first introduced in the text. For instance, “SoC” appears in line 180, but its full name is only explained in line 548. Please ensure this principle is applied consistently throughout the manuscript.

Re: We have carefully proofread all abbreviations appearing throughout the manuscript and corrected identified inconsistencies.

Review Comments from the Junior Reviewer:

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

This study introduces a low-cost, compact microscope combined with an AI-based screening system to enhance cervical cytology in resource-limited settings. By employing a two-stage approach--instance classification followed by an Att-Transformer model--the system achieved AUCs of 0.85-0.89 across multiple datasets, demonstrating its potential for efficient screening with minimal resources. The manuscript has improved significantly, with the authors effectively addressing my concerns and incorporating substantial experiments to support their conclusions. The reorganization enhances readability, and the supplementary materials provide valuable additional context. Overall, I appreciate the improvements and found the manuscript engaging to read.

[Response to the junior reviewer] We appreciate your valuable feedback, which has greatly improved our manuscript.

Minor points: junior reviewer

In Figure 1b, AB is labeled as the standard microscope in the plot. However, I noticed that the AB line is positioned in the lower-left section, which corresponds to the LR image acquired by the compact microscope. Please revise accordingly.

Re: In the revised manuscript, we have replaced the original figure with an updated version incorporating statistical annotations to enhance compliance with established data visualization standards.

Reviewer #1 (Remarks on code availability):

I did not assess the code as I am currently on a trip.

Re: The code website URL is <https://github/ShenghuaCheng/high-risk-Cervical-Precancerous-Screening>. Please feel free to contact us if you encounter any issues during your visit.

Reviewer #2 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Re: Thanks.

Reviewer #4 (Remarks to the Author):

The authors have fully addressed my concerns during the revision with additional experiments.

Re: Thanks.

Reviewer #4 (Remarks on code availability):

The code is fine with a readme file.

Re: Thanks.

Reviewer #5 (Remarks to the Author):

1 The revised manuscript states “Cytopathology-based screening, 72 which directly identifies cellular morphological changes, remains a necessary follow-up step after 73 a positive HPV test for definitive diagnosis of cervical precancer.” This is not correct. A positive HPV test is followed by colposcopy and biopsy with histopathologic diagnosis of cervical precancer. The incorrect statement added to the introduction implies that the new methodology can be used for definitive diagnosis. This is not consistent with accepted clinical practice.

[Response 5-1] We gratefully appreciate your insightful feedback. In response, we have refined the terminology in the revised manuscript to better reflect scholarly precision.

The updated text now reads: “*Cytopathology-based screening, which directly detects cellular morphological changes, continues to serve as a routine and effective screening approach.*”

2 The revised paper does provide AUC for datasets including ASCUS. However, performance suffers quite a bit and these results are presented in supplementary figures. It is not clear that the performance achieved with benign samples included is clinically relevant even if the method does outperform other computational approaches. The authors correctly state that it is important to detect HSIL lesions, but they must still be discriminated from ASC samples. “Our method achieved overall average AUC of 0.780 whereas AUCs of the comparison 334 methods are lower than 0.525 (supplementary Fig. 7).”

[Response 5-2] We deeply appreciate this important comment. While our method shows superior computational performance (AUC 0.78 vs. 0.52 in other methods), we agree that clinical utility requires further validation. Our evaluation of HSIL+LSIL+ASC detection (shown in Supplementary Fig. 10) showed that the majority of false negatives were ASC cases, while HSIL/LSIL specificity remained high. This confirms again our method reliably distinguishes true precancerous lesions (HSIL/LSIL) from benign/negative samples but has lower accuracy for ambiguous ASC cases. Overall, our DualCytoNet demonstrates feasibility and clinical significance for SIL population screening in resource-limited settings. However, for ASC subcategories, current performance remains insufficient for clinical application due to imaging resolution limitations. Future work will focus on hardware and methodological optimization to improve ASC case discrimination.

3 Finally, recommendations for co-testing with HPV and Pap generally recommend that patients are referred for colposcopy if the HPV test is positive even if the Pap is negative.

[Response 5-3] We fully concur with the reviewer's perspective. In well-resourced settings, current guidelines uniformly recommend colposcopy referral for all HPV-positive cases, regardless of Pap test results, when co-testing is implemented. In resource-limited settings, prioritizing confirmatory testing for high-risk groups may be necessary due to healthcare constraints.

Cytopathology and HPV testing serve complementary roles in cervical cancer screening: the former provides direct morphological assessment, while the latter identifies etiological factors. Despite the increasing adoption of HPV-based screening, cervical cytology remains a critical component of co-testing protocols and serves as the primary triage method for HPV-positive women in primary HPV screening programs.

This combined approach is particularly valuable given two key considerations: (1) the high prevalence of transient HPV infections in women of reproductive age, and (2) the fact that approximately 90% of HPV infections are cleared spontaneously within two years. Cytological evaluation thus plays an essential role in minimizing unnecessary colposcopy referrals and invasive biopsies, while maintaining screening efficacy.

The following papers were referred to:

[1] Perkins RB, Guido RS, Castle PE, Chelmow D, Einstein MH, Garcia F, Huh WK, Kim JJ, Moscicki AB, Nayar R, Saraiya M, Sawaya GF, Wentzensen N, Schiffman M; 2019 ASCCP Risk-based management consensus guidelines committee. 2019 ASCCP Risk-based management consensus guidelines for abnormal cervical cancer screening tests and cancer precursors. J Low Genit Tract Dis. 2020 Apr;24(2):102-131. PMID: 32243307.

- [2] Fontham ETH, Wolf AMD, Church TR, Etzioni R, Flowers CR, Herzig A, Guerra CE, Oeffinger KC, Shih YT, Walter LC, Kim JJ, Andrews KS, DeSantis CE, Fedewa SA, Manassaram-Baptiste D, Saslow D, Wender RC, Smith RA. Cervical cancer screening for individuals at average risk: 2020 guideline update from the American Cancer Society. *CA Cancer J Clin*. 2020 Sep;70(5):321-346. PMID: 32729638.
- [3] WHO guideline for screening and treatment of cervical pre-cancer lesions for cervical cancer prevention [Internet]. 2nd ed. Geneva: World Health Organization; 2021. PMID: 34314129.

Reviewer #6 (Remarks to the Author):

I find the responses to my previous comments satisfactory. My two remaining comments are rather regarding the response letter.

Firstly, in the response to my initial comments the authors state that HPV screening is appropriate for "lower-resource" regions. It is important and appropriate, as well as much better than cytology, in any resource setting. Cytology is no longer the primary option even in higher-resource settings. Secondly, the authors on page 30 of the response letter state "In combined screening, our model can effectively exclude a large number of cytopathology negative cases, thereby improving the specificity of co-testing. For instance, we can recommend patients who are HPV-positive and have the highest scores of our algorithm for biopsy, while those with relatively lower algorithm scores but still HPV-positive can be referred for endoscopic examination. "

We do not really use the term "endoscopy" for visualizing the cervix. Do you mean colposcopy? Biopsies are taken under colposcopy guidance, so the above sentence is very difficult to understand - group 1 will need referral for colposcopic examination if a biopsy is to result.

[Response to reviewer # 6] We appreciate your helpful suggestions.

Regarding your first point, we agree that HPV screening is indeed highly valuable in all resource settings, including high-resource regions. Current evidence and guidelines (e.g., WHO, USPSTF) increasingly support HPV-based screening as the primary strategy for cervical cancer prevention, regardless of regional resources. Cytology, while historically dominant, now plays a secondary or adjunct role even in well-resourced healthcare systems due to HPV testing's higher sensitivity and programmatic efficiency.

Regarding your second point, the original use of term "endoscopy" was an inadvertent terminological error, which we have now replaced with the corrected term "colposcopy". Our intended clinical recommendation is as follows:

For HPV-positive patients with negative cytological screening results, the risk of precancerous lesions is relatively low. In resource-limited settings, these individuals may reasonably defer colposcopy and biopsy procedures, and implement tiered management based on HPV genotype risk stratification and regional healthcare resource availability. However, patients testing positive on both HPV and our screening system should be prioritized for immediate biopsy to confirm diagnosis.